Research Methods Handbook

Miguel Centellas
University of Mississippi

June 4, 2016
Updated June 14, 2016

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License:
http://creativecommons.org/licenses/by-nc-sa/4.0/

Introduction

This handbook was written specifically for this course: a social science methods field school in Bolivia. As such, the offers a brief introduction to the kind of research methods appropriate and useful in this setting. The purpose of this handbook is to provide a basic overview of the social scientific methodology (both qualitative and quantitative) and help students apply this in "real world" contexts.

To do that, this handbook is also paired with some datasets pulled together both to help illustrate concepts and techniques, as well as to provide students with a database to use for exploratory research. The datasets are:

- A cross-sectional database of nearly 200 countries with 61 different indicators
- A time-series database of 19 Latin American countries across 31 years (1980-2010) with ten different variables
- Various electoral and census data for Bolivia

We will use those datasets in various ways (class exercises, homework assignments) during the course. But you can (and should!) also use them in developing your own research projects.

This handbook condenses (as much as possible) material from several other "methods" textbooks. A number of the topics covered here might seem too brief. And many of the more sophisticated approaches (such as multivariate regression, logistic regression, or factor analysis) aren't explored (although these almost never explored in most undergraduate textbooks). But this handbook was written mainly with the assumption that you *don't* have access to specialized statistical software (e.g. SPSS, Stata, SAS, R, etc.). Because of that, the quantitative techniques taught in this handbook will walk you through the actual *mathematics* involved, as well as how to use basic functions available in Microsoft Excel to do quantitative statistical analysis. A few major statistical tests that require special software are discussed (in Chapter 7), but mostly with an eye to explaining when and how to use them, and how to report them. In class, I offer specific walkthroughs and examples in SPSS and/or Stata, as available.

Mainly, I hope this handbook helps you become comfortable with the logic of "social" scientific research, which shares a common logic with the "natural" sciences. At the core, both types of scientists are committed to explaining the real world through empirical observation.

1 Basic Elements

For most of your undergraduate career so far, you have (hopefully) encountered some of the ideas of social science *research* as a *process* (as opposed to simply being exposed to the product of other peoples' research). This chapter presents a short crash course on the basic elements of what "doing" social science research entails. Some of the ideas may be familiar to you from other contexts (such as your "science" classes). Still, please follow closely because while social sciences are very much a branch of science, some of the distinctions between the "natural" sciences (biology, chemistry, physics, etc.) and the "social" sciences (anthropology, sociology, political science, economics, and history) have important implications for how we "do" social science research.

Most of you are probably familiar with the basic components of the **scientific method**, as encountered in any basic science course. The basic scientific method has the following "steps":

1. Ask a research question
2. Do some preliminary research
3. Develop a hypothesis
4. Collect data
5. Analyze the data
6. Write up your research

Although the scientific method is often described in a linear fashion, that's not always how it works in the real world. The following discussion summarizes some important components of the scientific method—including several frequently unstated ones, such as the underlying assumptions upon which scientific thinking is built upon.

But there are two important elements of scientific research that should be mentioned up front: First, science is **empirical**, a way of knowing the world based on observation. A phenomenon is "empirical" if it can be observed (either directly with my five senses, or by an instrument). This is an important boundary for science, which means a great many things—even important ones such as happiness or love—can't be studied by scientific means. At least not directly.

Second, science requires **replication**. Because science is based on empirical observation, its findings rest exclusively on that evidence. Other researchers should be able to replicate your research and come to the same conclusions. Over time, as replications confirming research findings build up, they take the form of **theories**, abstract explanations of reality (such as the theory of evolution or the theory of thermodynamics). The importance of replication in science has important consequences, both for how research is conducted and how and why we write our research findings in a particular way.

Social Scientific Thinking

As in all sciences (including the "natural" sciences), social scientific thinking is a way of thinking about reality. Rather than argue about what *should* be, social scientists tend to think about what *is*—and then seek to understand, explain, or predict based on *empirical* observation.

Chava Frankfort-Nachmias, David Nachmias, and Jack DeWaard (2015) identified six assumptions necessary for scientific inquiry:

1. Nature is orderly.
2. We can know nature.
3. All natural phenomena have natural causes.
4. Nothing is self-evident.
5. Knowledge is based on experience (empirical observation).
6. Knowledge is superior to ignorance.

Briefly, what this means is that we assume that we can understand the world through empirical observation, and we reject (as scientists) explanations that aren't based on empirical evidence. Certainly, there are other ways of "knowing." When we say that such forms of knowledge aren't "scientific" we aren't suggesting that such forms of knowledge have no value. Rather, we simply mean that such forms of knowledge don't rely on *empirical* observations or meet the other assumptions that underlie scientific thinking. It's also true that some of the most important questions may not be answered scientifically: "What is the purpose of life?" is a question that can't be answered with science (that's a question for philosophy or religion). But if we want to understand—empirically—how stars come into existence, why there's such diversity of animal life on earth, or how humanity evolved from hunters and gatherers to industrial societies, then science can offer answers. The scientific way of thinking assumes that, despite the chaotic nature of the universe, we can identify patterns (whether in the behavior of stars or voters) that can allow us to understand, explain, or predict other phenomena.

Implicit in the above list is a core ideal of the scientific process: **testability**. Above all, science is a way of thinking that involves testable claims. Because nothing is "self-evidence," all statements must be verified and checked against empirical evidence. This is why **hypotheses** play a central role in scientific research: Hypotheses are explicit statements about a *relationship* between two or more *variables* that can be tested by observation.

Although social scientific research is generally empirical, there are some types of social research that are non-empirical. Because this handbook focuses on social scientific research, we won't say much about those. But it's important to be aware of them both to more fully understand the broader parameters of social research and to have a clearer understanding of the distinction between empirical and non-empirical research.

Types of Social Research

We can distinguish different kinds of research along two dimensions: whether the research is *applied* or *abstract*, and whether the research is empirical or non-empirical. These mark differences both in terms of what the goals or purpose of the research is, as well as what kind of evidence is used to support it. The table below identifies four different types of research:

Table 1-1 Types of Research

	Applied	Abstract
Empirical	"Engineering" research	Theory-building
Non-empirical	Normative philosophy	Formal theory

Scholarship that seeks to describe or advocate for how the world "should be" is **normative philosophy**. This kind of research writing may build upon empirical observations and use these as

evidence in support of an argument, but it's not "empirical" in the sense that philosophical works are "testable." This kind of work is called *normative* research, since it deals with "moral" questions and making *subjective* value judgements. For example, research on human rights that proposes a code of conduct for how to treat refugees advances a moral position. Such arguments may be persuasive—and we may certainly agree with them—but they are not "scientific" in the sense that they can be tested and disproven. We are simply either convinced of them, or we aren't.

Another form of non-empirical research is **formal theory** (or sometimes "positive theory"). Unlike philosophy, however, this kind of research isn't normative (it doesn't "advocate" a moral position). A good analogy is to mathematics, which is also not a science. Formal theorists develop abstract models (often using mathematic or symbolic logic) about social behavior. This kind of research is most common in economics and political science, rather than in anthropology or sociology. Formal theory relies much more heavily on empirical research, since it uses established findings as the "assumptions" necessary to as the first parts of *deductive* "proofs" of the models. Because formal theory uses deduction to describe explicit relationships between concepts, it produces theories that could be tested empirically—although formal theory doesn't do this. For example, a number of models of political behavior are built on *rational choice* assumptions, and then expanded through formal mathematical "proofs" (similar to the kind of proofs done in geometry). Other researchers, however, could later come and test some of the findings of formal theory through empirical, scientific research.

Research that aims at developing theory, but does so through empirical testing, is called **theory-building research**. In principle, all scientific research contributes to testing, building, and refining theory. But theory-building research does so explicitly. Unlike formal theory, it develops explicit hypotheses and tests them by gathering and analyzing empirical evidence. And it does so (as much as possible) without a normative "agenda."[1] Generally, when we think of social scientific research, this is what comes to mind.

Finally, **engineering research** doesn't study phenomenon with detachment, but rather uses normative position as a guide. In other words, this kind of research has a clear "agenda" that is made explicit. This kind of research is common in public policy work that seeks to solve a specific problem, such as crime, poverty, or unemployment. Whereas theory-building research would view these issues with detachment, engineering research treats them as moral problems "to be solved." One example of this kind of research is the "electoral engineering" research that emerged in political science in the 1990s. Simultaneously building on—and contributing to—theories of electoral systems, many political scientists were designing electoral systems with specific goals in mind (improving political stability, reducing inter-ethnic violence, increasing the share of women and minorities in office, etc.). The key difference between engineering or policy research and normative philosophy, however, is that engineering research uses scientific procedures and relies on empirical evidence—just as a civil engineer uses the realities of physics (rather than imagination) when constructing a bridge.

All four types of research exist within the social science disciplines, but this handbook focuses on those that fall in the *empirical* (or "scientific") spectrum. Although the discussions about research

[1] There's a lot that can be said about objectivity and subjectivity in any kind of scientific research. Certainly, because we are human beings we always have normative interests in social questions. One way to address this is to "confront" our normative biases at various steps of the research process—especially at the research design stage. In general, however, if we make sure to make our research procedures transparent and adhere to the principles and procedures of scientific research, our research will be *empirical* and *normative* in nature

design and methodology is aimed at theory-building research, it also applies to engineering research. Even if your primary interest is in normative or formal-theoretic research, an understanding of empirical research is essential—if nothing else, it will help you understand how the "facts" you will use to build your normative-philosophical arguments or as underlying assumptions for formal models were developed (and which ones are "stronger" or more valid).

Research Puzzles

Although the basic scientific method always starts with "ask a question," good empirical research should always begin with a research puzzle. Thinking about a research puzzle makes it clear that a research question shouldn't just be something you don't know. "Who won the Crimean War?" is a question, and you might do research to find out that that France, Britain, Sardinia, and the Ottoman Empire won the war (Russia lost). But that's merely looking up historical facts; it's hardly a puzzle.

What we mean by "puzzle" is something that is either not clearly known (it's not self-evident) or there are multiple potential answers (some may even be mutually exclusive). "Who won the Crimean War?" is not a puzzle; but "Why did Russia lose the Crimean War?" is a puzzle. Even if the historical summary of the war suggests a clear reason for winning, that reason was derived by someone doing historical analysis. A research puzzle is therefore a question that will require not just research to uncover "facts," but also a significant amount of "analysis," weighing those facts to assemble a pattern that suggests an answer.

In the social science, we also think of "puzzles" as having a connection to theory. "Why did Russia lose the Crimean War?" is not just a question about that specific war. Instead, that question is linked to a range of broader questions, such as whether different regimes have different power capabilities, how balance of power dynamics shape foreign policy, whether structural conditions favor some countries, etc. In other words, a social science "puzzle" is simple one part of a larger set of questions that help us develop larger understandings about the nature of the world.

A research question should be stated clearly. Usually this can be done with a single sentence. Lisa Baglione (2011) offers some "starting words" for research questions:

- Why …?
- How …?
- To what extent …?
- Under what conditions …?

Notice that these are different from the more "journalistic" questions (who, what, where, when) that are mostly concerned with facts. One way to think about this is that answers to social scientific research questions lend themselves to sentences that link at least two concepts. The most basic form of an answer might be something like: "Because of x, y happened." This is discussed more clearly in the discussions about variables, relationships, and hypotheses. But first we should say something about units of analysis and observation.

Basic Components of Scientific Research

In addition to being driven by puzzle-type research questions, all scientific research shares the following basic components: clearly specified units of analysis and observation, an attention to variables, and clearly specified relationships between variables in the form of a hypothesis.

Units of Analysis & Observation

Any research problem should begin by identifying both the **unit of analysis** (the "thing" that will be studied, sometimes referred to as the **case**) and the **unit of observation** (the units for data collection). It's important to identify this before data is collected, since data is defined by a level of observation. For example, imagine we want to study presidential elections in any country. We might define each election as a unit of analysis; so we could study one single election or several. But we could observe the election in many ways. We could use national-level data, in which case our level of analysis and observation would be the same. But we could also look at smaller units: We could collect data for regions, states, municipalities, or other subnational divisions. Or we might conduct surveys of a representative sample of voters, and treat each individual voter as a unit of observation.

The key is that in our analysis, we may use data derived from *units of observations* to make conclusions about different *units of analysis*. When doing so, however, it's important to be aware of two potential problems: the ecological and individualistic fallacies.

Ecological Fallacy. The ecological fallacy is a term used to describe the problem of using *group*-level data to make inferences about *individual*-level characteristics. For example, if look at municipal-level data and find that poor municipalities are more likely to support a certain candidate, you can't jump to the conclusion that poor individuals are more likely to support that candidate in the same way. The reasons for this are complex, but a simple analogy works: If you knew the average grade for a course, could you accurately identify the grade for any individual student? Obviously not.

Individualistic Fallacy. The individualistic fallacy is the reverse: it describes using *individual*-level data to make inferences about *group*-level characteristics. Basically, you can't necessarily make claims about large groups from data taken by individuals—even a large representative group of individuals. For example, if you surveyed citizens in a country and found that they support democracy. Does this mean their government is a democracy? Maybe not. Certainly, many dictatorships have been put in place despite strong popular resistance. Similarly, many democracies exist even in societies with high authoritarian values.

Because researchers often use different levels for their units of analysis and units of observation, we do sometimes make inferences across different levels. The point isn't that one should *never* conduct this kind of research. But it does mean that you need to think very carefully about whether the kind of data collected and analyzed allows for conclusions to be made across the two levels. For example, the underlying problem with the example for individualist fallacy is that *regime type* and *popular attitude*s are very different conceptual categories. Sometimes, the kind of question we want to answer doesn't match up well with the kind of data we can collect. We can still proceed with our research, so long as we are aware of our limitations—and spell those out for our audience.

Variables

Any scientific study relies on gathering *data* about *variables*. Although we can think about any kind of evidence as a form of data (and certainly all data is evidence), the kind of data that we're talking about here is data that *measures* types, levels, or degrees of *variation* on some dimension.

One way to better understand variables is to distinguish them from **concepts** (abstract ideas). For example, imagine that we want to solve a research puzzle about why some countries are more "developed" than others. You may have an abstract idea of what is meant by a country's level of "development" and this might take cultural, economic, health, political, or other dimensions. But if you want to study "development" (whether as a process or as an endpoint), you'll need to find a way

to *measure* development. This involves a process of **operationalization**, the transformation of *concepts* into *variables*. This is a two-step process: First, you need to provide a clear definition of your concept. Second, you need to offer a specific way to measure your concept in a way that is variable.

It's important to remember that any measurement is merely an instrument. Although the measure should be conceptually **valid** (it should credibly measure what it means to measure), no variable is perfect. For example, "development" is certainly a complex (and *multidimensional*) concept. Even if we limited ourselves to an economic dimension (equating "development" with "wealth"), we don't have a prefect measure. How do we measure a country's level of "wealth"? Certainly, one way to do this is to use GDP per capita. But this is only an imperfect measure (why not some other economic indicator, like poverty rate or median household income?). In Chapter 3 we discuss different kinds (or "levels") of variables (nominal, ordinal, interval, and ratio). Although these are all different in important ways, they all share a similarity: By transforming concepts into variables, we move from abstract (ideas) to empirical (observable things). It's important to avoid **reification** (mistaking the variable for the abstract thing). GDP per capita isn't "wealth," any more than the racial or ethnic categories we may use are true representations of "race" (which itself is just a social construct).

In scientific research, we distinguish between different kinds of variables: *dependent*, *independent*, and *control* variables. Of these, the most important are dependent and independent variables; they're essential for *hypotheses*.

Dependent Variables. A dependent variable is, essentially, the *subject* of a research question. For example, if you're interested in learning why some countries have higher levels of development than others, the variable for "level of development" would be your dependent variable. In your research, you would collect data (or "take measurements") of this variable. You would then collect data on some other variable(s) to see if any variation in these affects your dependent variable—to see if the variation in it "depends" on variation in other variables.

Independent Variables. An independent variable is any variable that is not the subject of the research question, but rather a factor believed to be *associated* with the dependent variable. In the example about studying "level of development," the variable(s) believed to *affect* the dependent variable are the independent variable. For example, if you suspect that democracies tend to have higher levels of development, then you might include *regime type* (democracies and non-democracies) as an independent variable.

Control Variables. When trying to isolate the relationship between dependent and independent variables, it's important to think about introducing control variables. These are variables that are included and/or accounted for in a study (whether directly or indirectly, as a function of *research design*). Often, control variables are either suspected or known to be associated with the dependent variable. The reason they are included as *control* variables is to isolate the *independent* effect of the independent variable(s) and the dependent variables. For example, we might know that education is associated with GDP per capita, and want to control for the relationship between GDP per capita and regime type by accounting for differences in education. Other times, control variables are used to isolate other factors that we know muddy the relationship. For example, we may notice that many oil-rich authoritarian regimes have high GDP per capita. To measure the "true" relationship between regime type and GDP per capita, we should control for whether a country is a "petrostate."

How we use control variables varies by type of **research design**, type of methodology, and other factors. We will address this in more detail throughout this handbook.

Hypotheses

The hypothesis is the cornerstone of any social scientific study. According to Todd Donovan and Kenneth Hoover (2014), a hypothesis organizes a study, and should come at the *beginning* (not the end) of a study. A hypothesis is a clear, precise statement about a proposed relationship between two (or more) variables. In simplest terms: the hypothesis is a proposed "answer" to a research question. A hypothesis is also an *empirical* statement about a proposed relationship between the dependent and independent variables.

Although hypotheses can involve more than on independent variable, the most common form of hypothesis involves only one independent variable. The examples in this handbook will all involve only hypotheses involving one dependent variable and one independent variable.

Falsifiable. Because a hypothesis is an empirical statement, it is by definition *testable*. Another way to think about this is to say that a good hypothesis is "falsifiable." One of my favorite questions to ask at thesis or proposal presentations is: "How would you falsify your hypothesis?" If you correctly specify your hypothesis, the answer to that question should be obvious. If your hypothesis is "as x increases, y also increases," your hypothesis is falsified if in reality *either* "as x increases, y decreases" *or* if "as x increases, y stays the same" (this second formulation, that there is *no* relationship between the two variables, is formally known as the **null hypothesis**).

Correlation and Association. We most commonly think of a hypothesis as a statement about a *correlation* between the dependent and independent variables. That is, the two variables are related in such a way that the variation in one variable is reflected in the variation in the other. Symbolically, we might express this as:

$$y = f(x)$$

where the dependent variable (y) is a "function" of the independent variable (x). Mathematically, if we knew the value of x and the precise relationship (the mathematical property of the "function"), then you can calculate the value for y.

There are two basic types of correlations are:

- Positive correlation
- Negative (or "inverse") correlation

In a **positive correlation**, the values of the dependent and independent variables increase together (though they might increase at different rates). In other words, as x increases, y also increases. In a negative or **inverse correlation**, the two variables move in opposite directions: as x increases, y decreases (or vice versa).

The term "correlation" is most appropriate for certain kinds of variables—specifically, those that have precise *mathematical* properties. Some variable measures, as we will see later, don't have mathematical properties; then it's more appropriate to speak about **association**, rather than correlation. For those kind of association, the relationship for a positive association takes the form "if x, then y." And a negative association takes the form "if x, then *not* y."

Causation. It's very important to distinguish between correlation (or association) and *causation*. Demonstrating correlation only shows that two variables move together in some particular way; it

doesn't state which one causes a variation in the other. Always remember that the decision to call one variable "dependent" is often an arbitrary one.

If you claim that the observed changes in your independent variable *causes* the observed changes in your dependent variable, then you're claiming something beyond correlation. Symbolically, a causal relationship can be expressed like this:

$$x \rightarrow y$$

In terms of association, a causal relationship goes beyond simply observing that "if x, then y" to claiming that "because of x, then y."

While correlational properties can be measured or observed, causal relationships are only *inferred*. For example, there's a well-established association between democracy and wealth: in general, democratic countries are richer than non-democratic ones. But which is the cause, and which is the effect? Do democratic regimes become wealthier, faster than non-democracies? Or do countries become democratic once they achieve a certain level of wealth? This chicken-or-egg question has puzzled many researchers.

It's important to remember this because correlations can often be products of random chance, or even simple artefacts of the way variables are constructed (we call this **spurious correlation**). More importantly, correlations may also be a result of the reality that some *other* variable is actually the cause of the variation in both variables (both are "symptoms" of some of other factor).

There are three basic requirements to establish causation:

- There is an observable correlation or association between x and y.
- Temporality: If x causes y, then x must precede y in time. (My yelling "Ow!" doesn't cause the hammer to fall on my foot.)
- Other possible causes have been ruled out.

Notice that correlation is only one of three logic requirements to establish causation. Temporality is sometimes difficult to disentangle, and most simple statistical research designs don't handle this well. But the third requirement is the most difficult. Particularly in the more "messy" social sciences, it is often impossible to rule out *every* possible alternative cause. This is why we don't claim to *prove* any of our hypotheses or theories; the best we can hope for is a degree of *confidence* in our findings.

The Role of Theory

Social scientific research should be both guided by and hope to contribute to theory. One reason why theory is important is because it helps us develop *causal* arguments. Puzzle-based research is *theory-building* because it develops, tests, and refines *causal* explanations that go beyond simply describing what happened (Russia lost the Crimean War), but try to develop clear explanations for *why* something happened (why did Russia lose the war?). Even if your main interest is simply curiosity about the Crimean War, and you don't see yourself as "advancing theory," an empirical puzzle-based research contributes to theory, because answering that question contributes to our understanding of *other* cases beyond the specific one. Understanding why Russia lost the Crimean War may help us under why countries lost wars more broadly, or why alliances form to maintain balance of power, or other issues. Understanding why Russia lost the Crimean War should help us understand other, similar phenomena.

Theories are not merely "hunches," but rather systems for organizing reality. Without theory, the world wouldn't make sense to us, and would seem like a series of random events. One way to think about theories is to think of them as "grand" hypotheses. Like hypotheses, theories describe links between concepts. Unlike hypotheses, however, theories link *concepts* rather than *variables* and their sweep is much broader. You might hypothesize that Russia lost the Crimean War because of poor leadership. But this could be converted into a theory: Countries with poor leaders lose wars. The hypothesis is about a *particular* event; the theory is *universal* because it applies to all cases imaginable.

While hypotheses are the cornerstones of any scientific study, theories are the foundations for the whole practice of science. Hoover and Donovan (2014, 33) identify four important uses of theory:

- Provide patterns for interpreting data.
- Supply frameworks that give concepts and variables significance (or "meaning").
- Link different studies together.
- Allow us to interpret our findings.

Not surprisingly, any research study needs to be placed within a "theoretical framework." This is in large part the purpose of the literature review. A good literature review is more than just a summary of important works on your *topic*. A good literature review provides the theoretical foundation that sets up the rest of your research project—including (and especially!) the hypothesis.

Fundamentally, theories a good theory is **parsimonious** (many call this "elegant"). Parsimony is the principle of simplicity, of being able to explain or predict the most with the least amount. This is important, because we don't strive for theories that explain *everything*—or even theories that can explain 100% of some specific phenomenon. Many things explain the French Revolution, for example, but a good theory is one that can do a good job of explaining that event with the fewest amount of variables.

Perhaps the easiest way to understand this is to actually think about some "big" theories. Although there are many, many social scientific theories, these can be merged into larger camps, approaches, or even paradigms. Lisa Baglione (2016, 60-61) identified four "generic" types of theories: interest-based, institutional, identity-based (or "sociocultural"), and economic (or "structural"). It may help to see how we can apply each of these generic theories to a simple question: What explains (or "causes") why some countries are democracies, and others are not?

Interest-Based Theories

Interest-based theories focus on the decisions made by actors (usually individuals, but can also be groups or organizations treated as "single actors"). Perhaps the most common is **rational choice theory**, which is a theory of social behavior that assumes that actors make "rational" choices based on a cost/benefit calculus.

Interest-based theories of democracy might argue that democracies emerge (and then endure) because all the relevant actors have decided to engage in collective decision-making because the costs of refusing to play outweigh any sacrifices necessary to play and/or the benefits of playing the democratic game outweigh any losses. This tradition helps explain democratic "pacts" between rival elites (which includes leaders of social movements, a common way of understanding democratic transitions in the 1980s. In particular, rational choice theories often involve game metaphors: games involve actors (players) who make strategic decisions based on how the other players will act. In this tradition, Juan Linz and Alfred Stepan (1996, 5) once declared that democracies were consolidated when they became "the only game in town" because actors were no longer willing to walk away from the table and play a different game (such as the "coup game").

Institutional Theories

Institutional theories focus on the "rules"—or institutions—that shape political life as deciding the most important factors. Institutions are, broadly speaking, the sets of formal or informal norms that shape behavior. Although more formalistic legal studies were important in the study of politics a century ago and earlier, that kind of legalistic studies fell out of favor during the behavioral revolution (which, among other things, put individual actors at the center of social explanations). But by the 1980s a **"new" institutionalism** had begun to emerge that once again put emphasis on institutions—but this time placing equal emphasis on formal *and informal* institutions that shape politics. Formal institutions include things like executives, legislatures, courts, and the laws that dictate their relationships. But they can also include less formal institutions, like political parties or interest group associations. In fact, some countries only have "informal" institutions: Great Britain has *no* written constitution; all of its governing institutions in some sense are "informal" (they are norms that are followed, which is what really matters).

Institutional theories about democracy—or at least democratic stability—became very common in during the 1990s. Some argued that presidential systems were inherently unstable, compared to parliamentary systems. Juan Linz (1994) made the argument that presidential institutions, with their separation of powers and conflicting legitimacy (both the executive and the legislature are popularly elected, so can each claim a "true" democratic mandate), were toxic and helped explain why no presidential democracy (other than the US) had endured more than a two or three decades. Reforming institutions also became an important area of *practical* ("engineering") research, including efforts by political scientists to (re)design new institutions to reform or strengthen democracy in various ways by studying whether certain electoral systems were more likely to better represent minorities, or government stability, etc.

Sociocultural Theories

The category of theory Baglione referred to as "ideas-based" is something of a catch-all for actor-centered explanations that are *not* interest-based or rational choice explanations. In other words, rather than operating on the basis of their *material* interests, "ideas-based" theories argue that individuals make decisions based on their inner beliefs. This can come from an ideology, but it can also come from culture and cultural values.

Sociocultural explanations of politics aren't very popular today, mainly because they have a history of reducing cultures to caricatures. For example, as late as the 1950s, many believed that democracy was incompatible with cultures that weren't Protestant. After all, beyond a handful of exceptional cases, the only democracies in the 1950s were in predominantly Protestant countries (northern Europe, the US and Canada, and a few others). Many argued that predominantly Catholic countries were incompatible with democracy—at least until they became less religious and more secular. And yet the 1970s and 1980s saw a massive "third wave" of democratization across most of the Catholic world (southern Europe and Latin America). Many who today argue that Islam is "incompatible" with democracy are likely making the same mistake.

But in many ways culture (and ideologies more generally) do matter and clearly influence individual behaviors. After all, we all grow up and are **socialized** to believe in many things, which we then take for granted. Often, we make decisions without really going through complex calculations to maximize our interests, but rather simply because we believe it's the way we are "supposed" to behave.

Economic or "Structural" Theories

Structural theories place large systems—generally economic ones—at the center of explanations for how the world works. "Structuralists" see human behavior as shaped by external forces (systems or "structures") over which they have limited control. Perhaps the most well-known structural theory is Marxism. Although the term is often used with an *ideological* connotation, in social science Marxism is often associated with a form of economic structuralism. After all, Marx developed his belief in the inevitability of a future (world) socialist revolution (the basis of Marxism as an ideology) on his analysis of world history: The evidence he gathered convinced him that every society was shaped by class conflict, which was in turn determined by the "mode of production" (economic forces); when those economic forces changed, the old status quo fell apart and new class conflicts emerged. In other words, economic forces not only shaped society, they also shaped its political. Any time someone explains politics with the slogan "it's the economy, stupid" they're engaging in Marxist, structural analysis.

Even many anti-communists have adopted "Marxist" understandings of reality to explain modern society (and sometimes to advocate for policies to shape society). Proponents of **modernization theory** argued that economic transformations would lead to democratization. They argued that as countries developed economically (they became wealthier, more industrialized) these economic changes would transform their societies (they "modernize") which in turn would set the foundation for democratic politics. During the Cold War, some even justified military regimes as necessary to provide the stability needed for the economic reforms that would drive modernization—which would eventually lead to democratic transitions. Other kinds of modernization theories analyze how changes in economic structures are related to social, political, or cultural changes.

Agency *vs*. Structure

Another way to think about differences between theories is whether they emphasize the role of *agency* (the ability of individuals to make their own free choices) or *structure* (the role that external factors play in shaping individual choices. In a simple sense, this is a philosophical debate between free will and fate or determinism. Do social actors make (and remake) the world as they wish? Or do social actors simply play out their "roles" because of structural constraints? Of course, the real world is too complicated for any either extreme to be universally "true." But remember that an important goal of theory is to be *parsimonious* (or "simple"). We adopt an emphasis on agency or structure as a sort of heuristic device in order to try to explain a complex event by breaking it down into a handful of related concepts.

The four "big" theoretical perspectives described above can also be sorted into whether they emphasize agency or structure. The one exception is the larger "ideas-based" group of theories Baglione described. I renamed it "sociocultural theories" to distinguish the role of ideology or culture from a different set of ideas-based theories that emphasize *psychological* factors. These are actor-centered approaches (like rational choice) but don't assume that actors behave "rationally" (follow their best "interests").

2 Research Design

Research design is a critical component of any research project. The way we carry out a research project has important consequences for the validity of our findings. It's important to spend time at the early stage of a project—even *before* starting to work on a literature review—thinking about *how* the research will proceed. This means more than selecting secondary or even primary sources of data. Rather, research design means thinking carefully about how to structure the logic of inquiry, what cases to select, what kind of data to collect, and what type of analysis to perform.

Thinking about research design involves thinking about three different, but related issues:

- How many cases will be included in the study?
- Will the study look at changes over time, or treat the case(s) as essentially "static"?
- Will you use a qualitative or quantitative approach (or some mix of both)?

The answer each question largely depends on the kind of data available. If data is only available for a few cases, then a large-N study is simply not possible. If quantitative evidence isn't available (for certain cases and/or time periods), then you may have to rely on qualitative evidence. Then again, perhaps some questions are best answered qualitatively. The question itself also affects the kind of research design that is better suited to answering it. There's no "right" research design for any given situation—but there are "better" choices you can make.

It helps to remember that research designs should be *flexible*. For various reasons, you may need to revisit it once your project is underway. This may mean changing the number of cases (or even swapping out cases), changing from a cross-sectional to a time-series design, or moving between qualitative or quantitative orientations. Flexibility doesn't mean to simply use whatever evidence is available willy-nilly. Instead, flexibility means being able to adopt another type of research design. In order to be flexible, however, you must first be familiar with the underlying basic logic of *scientific* research.

Basic Research Designs

The purpose of a research design is to help us test whether there does in fact exist a relationship between the two variables as specified in our hypothesis. As in all scientific studies, this involves a process of seeking to reduce alternative explanations. After all, our two variables may be related for reasons that have nothing to do with our hypothesis.

W. Phillips Shively (2011) identified three types of basic research designs: true experiments, natural experiments, and designs without a control group.

True Experiments

When you think of the *scientific method*, you probably think about laboratory experiments. Not surprisingly, experimental designs remain the "gold standard" in the sciences—including the social sciences. This is because experiments allow researchers (in theory) perfect control over research conditions, which allows them to isolate the effects of an independent variable.

An experimental research design has the following steps:

1. Assign subjects at random to both test and control groups.
2. Measure the dependent variable for both groups.
3. Administer the independent variable to the test group.
4. Measure the dependent variable again for both groups.
5. If the dependent variable changed for the test group relative to the control group, ascribe this as an effect of the independent variable.

A key underlying assumption of the experimental method is that both the test and control groups are similar in all relevant aspects. This is key for control, since there should be no differences between the groups because any difference would introduce yet another variable, which means we can't be certain that the independent variable (and not this other difference) is what explains our dependent variable.

Researchers attempt to ensure that test and control groups are similar through *random selection* of cases. Even so, whenever possible, it's important to check to make sure that the selected groups are in fact similar. There are statistical ways to check to see whether two groups, which we will discuss later. But a good rule of thumb is to always keep asking whether there's any reason to think the cases selected are appropriately *representative* of the larger population, or at least (in an experimental design) similar enough to each other.

Although experiments are becoming more common in many areas of social science research, it may be obvious that many research areas can't—either for ethical or practical considerations—be subjected to controlled experimentation. For example, we can't randomly assign countries to control and test groups, and then subject one group to famine, civil war, or authoritarianism just to see what happens.

Natural Experiments

When true experiments aren't an option, researchers can approximate the conditions if they can find cases that allow them to look at a "natural" experiment.

A natural experiment design has the following steps:

1. Measure the dependent variable for both groups *before* one of the groups is exposed to the independent variable.
2. Observe that the independent variable
3. Measure the dependent variable again for both groups.
4. If the dependent variable changed for the group exposed to the independent variable relative to the "control" (unexposed) group, ascribe this as an effect of the independent variable.

Notice that the only significant difference between "natural" and "true" experiments is that in natural experiments, the researcher has no control over the introduction of the independent variable. Of course, this also means he/she also doesn't have any control over which cases fall into which group—and therefore only a limited ability to ensure that the two groups are in most other ways similar. Still, with careful and thoughtful case selection, a researcher can select cases to maximize the ability to make good inferences.

One classic example of a natural experiment is Jared Diamond's (2011) study of the differences between Haiti and the Dominican Republic, two countries that share the island of Hispaniola.

Despite sharing not only an island, but a common historical experience with colonialism, the two countries diverged in the 1800s. Today, Haiti is the poorest country in the hemisphere, while the Dominican Republic ranks on most dimensions as an average Latin American country.

A natural experiment still requires measurement of both test and control group(s). Diamond's natural experiment of the two Hispaniola republics depends on the fact that he was able to observe the historical trajectories of both countries for several centuries using the historical record. This allowed him to identify moments when the two countries diverged in other areas (forms of government, agricultural patterns, demographics, etc.) that explain their diverging economic development trajectories.

Sometimes, however, we may find two cases that potential represent a natural experiment, but for whom no pre-measurement is possible. This variation looks like:

1. Measure the dependent variable for both groups *after* one of the groups is exposed to the independent variable.
2. If the dependent variable is different between the two groups, ascribe this as an effect of the independent variable.

While this design is clearly not as strong, sometimes it's the best we can do. In that case, it's important to be *explicit* about the limitations of this type of design—as well as the steps taken to ensure (as much as possible) that the cases/groups were in fact similar before either was exposed to the independent variable.

Designs Without a Control Group

Yet another basic type of research design is one that doesn't include a control group at all. It looks like this:

1. Measure the dependent variable.
2. Observe that the independent variable occurs.
3. Measure the dependent variable again.
4. If the dependent variable changed, ascribe this as an effect of the independent variable.

This design *requires* that pre-intervention measurements are available. Essentially, this type of research design treats the test group prior to the introduction of the independent variable as the control group. If nothing other than the independent variable changed, then any change in the dependent variable is logically attributed to the independent variable.

The Number of Cases

The number of cases (units of observation) is an important element of research design. Choosing the appropriate cases—and their number—depends both on the research question and the kind of evidence (data) that is available. Many questions can be answered by many different kinds of research designs; there is no "right" choice of cases. However, it's important to keep in mind that the number of cases has implications for how you treat time, as well as whether you pursue a qualitative or quantitative approach.

There are three types of research designs based on the number of cases: large-N studies, which look at a large number of cases ("N" stands for "number of cases"); comparative studies, which look at a small selection of cases (often as few as two, but no more than a small handful); and case studies, which focus on a single case. In all three, *how* the cases are selected is very important, but perhaps most so as the number of cases gets smaller.

Case Studies

In some ways, a case study—an analysis of a single case—is the simplest type of research design. However, this doesn't mean that it's the easiest. Instead, case studies require as much (if not more!) careful thought. A case study is essentially a design without a control group. This means that a case must be studied *longitudinally*—that is, over a suitably period of time. This is true regardless of whether the case study is approached as a qualitative or quantitative study. Finally, this also means that the selection of the case for a case study is critically important, and shouldn't be made randomly.

One important thing to remember is that in picking case studies, a researcher must *already* know the outcome of the dependent variable. A case study seeks to explain *why* or *how* the outcome happened. For example, suppose we pick Mexico as a case to study the consolidation of a dominant single-party regime in the aftermath of a social revolution. The rise of Mexico's PRI is taken as a *social fact*, not an outcome to be "demonstrated."

Two basic strategies for selecting potential cases for a case study are to pick either "outlier" or "typical" cases. This means, of course, that a researcher must be familiar not only with the cases they want to study, but also the broader set of patterns found among the population of interest. Even if you come to a project with a specific case already in mind (because of prior familiarity or because of convenience or for any other reason), you should be able to identify whether the case is an outlier or a typical case. If a case is not quite either, then you should either select a different case or a different research design. This is because each type of case study has different strengths that lend themselves to different purposes.

Outlier Cases. "Outliers" are cases that don't match patterns found among other similar cases or in ways predicted by theory. Studies of outlier cases are useful for *testing* theory. While a single deviant case might not "disprove" an established theory all on its own, it certainly reduces the strength of that theory. Additionally, a study of an outlier case may show that another factor is also important in explaining a phenomenon. For example, there's a strong relationship between a country's level of wealth and its health indicators. Yet despite being a relatively poor country, Cuba has health indicators similar to that of very wealthy countries. This suggests that although a country's wealth is a strong predictor of its health, other factors also matter. In some cases, the study of outlier cases may reveal that an outlier really isn't an outlier on close inspection.

Typical Cases. "Typical" cases cases match broader patterns or theoretical expectations. While studies of typical cases don't do much to test theory, they can help *explain* the mechanisms that underlie a theory. This is because while large-N analysis is stronger at demonstrating *correlations* between variables, it isn't very useful for demonstrating *causality*. For example, knowing that health and wealth are correlated tells us little about the direction of that relationship, or *how* wealth or health affects the other. One way to do this through *process tracing*, a technique that focuses on the specific mechanisms that link two or more events, and carefully analyzing their sequencing.

Comparative Studies

Studies of two or more cases are commonly referred to as "comparative studies." A good way to start a comparative study is to begin by selecting an "outlier" or "typical" case, just like in a single-case study, and then find an appropriate second case. Two basic strategies for selecting cases for a comparative study identified by Henry Teune and Adam Przeworski (1970) are the "most-similar" and "most-different" research designs. As with case studies, a researcher needs to be familiar with the individual cases, as well as broader patterns. Selecting cases for a comparative design requires additional attention, since the cases must be convincingly similar/different from each other.

Most-Similar Systems (MSS) Designs. MSS research designs closely resemble a natural experiment. The logic of this design works this way: If two cases closely resemble each other in most ways, but differ in some important outcome (dependent variable), then there must be some other important difference (independent variable) that explains why the two cases diverge on the dependent variable. Essentially, all the ways the two cases are similar cancel each other out, and we are left with the differences in the dependent and independent variables.

Imagine two cases that are similar in various ways (X_n), but have different outcomes (Y_1 and Y_0).

Case 1: $X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot X_5 \cdot A \rightarrow Y_1$

Case 2: $X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot X_5 \cdot B \rightarrow Y_0$

Logic suggests that since similarities can explain different outcomes, there must exist at least one other difference between the two cases. Looking carefully at the two cases, we find that they have different measures (A and B) on one variable.

One simple strategy for selecting cases for MSS designs is to find cases that diverge on the dependent variable, then identify a "most similar" pair of cases. For example, if you wanted to understand what causes social revolutions in the twentieth century, you might select one classic example of social revolution (Bolivia) and a similar country (Peru) that did not experience a social revolution in the twentieth century.

It's tempting to think of a single-case study as a "most similar" design, particularly if we carefully divide one "case" into two observations. But because the case moves forward through time, too many other changes also occur that make it difficult to isolate independent variables.

Most-Different Systems (MDS) Designs. MDS research designs are the inverse, but use the same underlying logic: If two cases are in most ways different from each other, but are similar on some important outcome (dependent variable), there must be some other similarity (independent variable) that explains this convergence. One simple strategy for selecting cases for MDS designs is to find cases that match up on the dependent variable, then identify a "most different" pair of cases. For example, if you wanted to study of pan-regional populist movements, you might select two countries that experienced such movements, but came from different regions: Peru (*aprismo*) and Egypt (*Nasserism*).

Combined MSS and MDS Research Designs. There are many ways to combine MSS and MDS research designs. One possibility is to first pick a MSS design, and then add a third case that pairs up with one of those cases as a MDS comparison. For example, in our MSS example above we picked Peru and Bolivia as similar cases. We might then look for another country that also had a social revolution, but was very different from Bolivia. Alternatively, we might look for another country that also did not have a social revolution, but was very different from Peru. A second possibility is to start with a MDS design, and then add a third case that pairs up with one of those cases as a MSS comparison. In both cases, the logic would be one of *triangulation*: combining both MSS and MDS designs allows a researcher to cancel out several factors and zero in on the most important independent variables.

Large-N Studies

Any study involving more than a handful of cases (or observations) can be considered a large-N study. Large-N studies have important advantages because they come closest to approximating the

ideal of experimental design. In fact, experimental designs are stronger the larger their test and control groups, since larger groups are more likely to be *representative*, making findings more valid and the conclusions more generalizable.

Usually, large-N studies look at a *sample* of a larger *population*. This is particularly true when the study looks at individuals, rather than aggregates (cities, regions, countries). It's tempting to think that a study of all the world's countries is a study of the universe of countries, but this is rarely the case. Beyond the question of what counts as a "country" (are Taiwan, Somaliland, or Puerto Rico "countries"?) lies the reality that we often don't have full data on all countries, which means that such studies invariable exclude some cases. Therefore, we should think about all large-N studies as studies of "samples."

This means that large-N studies must be concerned with whether the cases included in the study (the sample) are representative of the larger "population" (the universe of all possible cases). Later in this handbook, we'll look at statistical ways to test whether a sample is representative. But you should at least think about the cases that are excluded and consider whether they share any characteristics that need to be addressed. Sometimes cases are excluded simply because data isn't available for some of them. But the lack of data may also be correlated with some other factors (level of development, type of government, etc.) that might be important to consider.

Finally, because cross-sectional studies look at a large number of cases, the ability to offer significant detail on any of the cases is diminished. This means that large-N studies tend to be more *quantitative* in orientation; even when some of the variables are clearly *qualitative* in nature, they are treated as quantitative in the analysis.

There are two basic types of large-N studies: cross-sectional and time series studies. The logic of both is essentially the same, but there are some important differences. Later in this handbook, we'll look at some quantitative techniques used to measure relationships in both types of studies.

Cross-Sectional Studies. Studies that look at a many cases (whether individuals or aggregates) using a "snapshot" of a single point in time are considered cross-sectional studies. The purpose of a cross-sectional study is to identify broad patterns of relationships between variables.

It's important to remember that cross-sectional studies treat all observations as "simultaneous," even if that's not the case. For example, if you were comparing the voter turnout in countries, you might use the most recent election—even if the recorded observations would vary by several years across the countries. You'll often see that cross-sectional studies use "most recent" or "circa year X" as the time reference. The important thing is that each case is observed only once (and that the measurements are "reasonably" in the same time frame).

Time-Series Studies. Unlike cross-sectional studies, time-series studies include a *temporal* dimension of analysis. They also consider one *case*, divided into a large number of *observations*, but analyzed in a more formal and *quantitative* way. A time-series study of economic development in Bolivia would differ from the more *qualitative* narrative type of analysis of a traditional single case study because it would divide the case into a large number of observations (such as by years, quarters, or months) and provide discrete measurements of each time unit.

The simplest form of time-series analysis is a bivariate analysis that would simply treat time as the independent variable (x) and see whether time was meaningfully correlated with an increase or decrease in the dependent variable (y). This can be done with simple linear regression and

correlation (explained in Chapter X). In some cases, time can be introduced in a three-variable model using partial correlation (explained in Chapter X).

Panel Studies. Studies that combine cross-sectional and time-series analysis are called panel studies. The simplest form of a panel study involves a collection of cases and measuring each one twice, for a series of before/after comparisons. These can be analyzed with two-sample difference of means tests, explained later in this handbook (see Chapter X). But more sophisticated panel studies involve collecting data from multiple points in time for each observation. These require much more care than the simpler cross-sectional and time-series designs. While this handbook doesn't cover these, they can be handled with most statistical software packages.

Mixed Designs

Because there is no single "perfect" research design, it's useful to combine more different kinds of research designs into a single research project. For example, a large-N cross-sectional study can be used to identify an "outlier" or a "typical" case for a qualitative case study. Or you can combine a cross-sectional large-N design with a time-series large-N study of a single case. You can also combine large-N and comparative studies, or combine two types of comparative studies (MSS and MDS) with a more detailed case study of one of the cases. Thinking creatively, you can mix different research designs in ways that strengthen your ability to answer your research question.

One special kind of mixed design is a *disaggregated* case study. For example: Imagine you wanted to do a case study of Chile's most recent election. If you didn't want to add a comparison case, but wanted to increase the number of observations, you could do this by adding studies of subunits. These could be regions, cities, or even individuals (for example, with a survey or a series of interviews). If the subunits were few in number, you could select some for either an MSS or MDS comparison. If the subunits were of sufficient number, you could treat this as a large-N analysis to support the analysis made in the country-level case study. For example, if you have data for Chile's 346 communes (counties), you could do a large-N analysis of election patterns. You could also do the same with survey data (either your own or publicly available survey data, such as that available from LAPOP). Or you could select two or three of Chile's 15 regions to provide additional detail and evidence. In this case, the *unit of analysis* (country) and the *unit of observation* (region, commune, or individual) are different. It's useful to remember that any social aggregate (a country, a political party, a school) can be disaggregated to lower-level units of observation.

Dealing with Time

All research studies must pay attention to time. Some research designs do so explicitly: cross-sectional studies look at one snapshot in time; time-series studies use time as one of the variables in the analysis. But even here, time needs to be *explicitly* discussed. A cross-sectional study should be clear about *when* the single "snapshot" in time comes from. Sometimes, it's as easy as simply saying that you will use the "most recent" data available—but even then you should be cautious. Cross-sectional data may come from across different years; every country has its own electoral schedule, for example. Time is also important when working with cases—whether as individual case studies or comparative studies of a handful of cases. After all, a study of "France" isn't as clear a study of "France in the postwar era."

Time in Case Studies

Because case studies are studied *longitudinally*, they are not momentary "snapshots" in time (as in cross-sectional studies). But the "time frame" for a case study should be clearly and explicitly defined. This means that a case study should have clear starting and ending points. If you are

studying Mexico during the Mexican Revolution, you should clearly define when this period began, and when it ended. Keep in mind that *you* define these periods, based on what you think is best for answering your question. The important thing in the example isn't to "correctly" identify the start and end of the Mexican Revolution, but rather to clearly state for your reader (and yourself) what you will and will not analyze in your research. Certainly, history constantly moves forward, so what happened before your time frame and what came after may be "important" and may merit some discussion. But they will not be included in your *analysis*.

Time in Comparative Studies

You can think of each case in a comparative study as a case study. All of the advice about time as related to individual case studies applies. But an important issue to keep in mind when it comes to comparative studies is that the two (or more) cases can be *asynchronous*. That is, the cases used in a comparative study can come from different time periods. The important thing is that the cases are either "most similar" or "most different" in useful ways. For example, Theda Skocpol's famous *States and Social Revolutions* (1979) compared the French, Russian, and Chinese revolutions. Thinking creatively about how select cases for comparison is important.

One other way to select cases for comparative studies is to break up a single case study into two or more specific "cases." This means more than simply describing the two cases as "before" and "after" some important event. If your research question is to explain why the French Revolution happened, this should be a single case study analyzed *longitudinally* by tracing the process over time. But if your research question seeks to understand the foreign policy orientations of different regimes, then a study of monarchist France and republican France could be an interesting comparison, since the two cases are otherwise "most similar" but with only different regime types. Breaking up a single case into multiple cases is a common "most similar" comparative strategy. Any study comparing two presidential administrations or two elections in the same country is essentially a "most similar" research design. Often, these are done implicitly. But there is tremendous advantage to doing so explicitly.

Time in Cross-Sectional Large-N Studies

Cross-sectional studies are explicitly studies of "snapshots" in time. The logic of cross-sectional analysis assumes that all the units of observation (the cases) are *synchronous*. This means great care should be given to making sure that all the cases are from "similar" time periods. Usually this means from the same year (or as close to that as possible), but this is a little more complicated that it seems.

One common form of cross-sectional analysis is to compare a large number of countries. For example, imagine that we want to study the relationship between wealth and health. We could use GDP per capita as a measure of wealth and infant mortality as a measure of health. Data for both indicators is readily available from various sources, including the World Bank Development Indicators. Imagine that we pick 2010 as our *reference* (or "snapshot") year. We might find that some countries are missing data for one or both indicators for that year. Should we simply drop them from the analysis? We could, but that has two potential side effects: it reduces the number of observations (our "N"), which has consequences for statistical analysis, and it could introduce bias if the cases with missing data share some other factors that make them different from the rest of the population.

One solution is to look at the years before and after for missing observations, and see if data is available for those years. The problem with this approach is that in this case we would be comparing data from different years, which may introduce other forms of statistical bias.

Another solution is to take the average for each country for some period centered around 2010 (say, 2005-2015). This also ensures that the data for the two variables are from the same reference point (so that you're not comparing 2011 GDP per capita with 2008 infant mortality, or similar discrepancies, for many observations). This solution has the added benefit of account for *regression to the mean*. For a number of reasons, data might fluctuate around the "true" value. If you take a single measure, you don't know whether that measure was an outlier (abnormally high or low). If the number is assumed to be relatively consistent, taking the mean of several measures is more likely to produce the "true" value. But this also isn't a perfect solution, since some countries may have only one or two data points, making their averages less reliable than those with ten data points. And some variables are not steady, but changing—and in different ways for different cases. No solution is perfect, and picking one will depend on a careful look at the data and thinking through the potential costs and benefits of each choice. In any case, your process for selecting the cases—and your justifications for that process—should be explicitly presented to readers.

Yet another way to select cases for cross-sectional analysis is to select the "most recent" data for each case. This is clearly appropriate for studies in which one or more variables in question is made up of discrete observations. For example, elections do not happen every year. So a cross-sectional study of voter turnout shouldn't limit itself to voter turnout across a specific reference year. You could calculate averages for some time period, but voter turnouts might fluctuate based on the idiosyncrasies of individual elections. Using the most recent election for each country is perfectly acceptable. However, it's important that any additional variables should match up with the year of the election. In other words, if you are doing a cross-sectional study that looks at "most recent" elections, you need to be sure that each country's data is matched up with that reference point.

There is room to think creatively in selecting cases for cross-sectional studies. For example, imagine that you wanted to understand factors that contribute to military coups in twentieth century Latin America. You could identify each of the military coups that took place in the region and treat each one as a "case" (and, yes, this means you could have multiple "cases" from a single country). You could then collect data on the time period of the coup and build a dataset for use in statistical cross-sectional analysis.

Time in Time-Series Large-N Studies

It may seem obvious that time plays a role in time-series analysis. But it's still worth being explicit about it. Because time-series studies are essentially case studies *disaggregated* into a large number of "moments," it's important to do two things: identify what counts as a "moment," and identify the study's time frame.

The concerns about identifying "moments" is similar to those for cross-sectional analysis, except that the logic of time-series requires that all the moments be identical. That is, you should decide what unit of time you will use (years, quarters, months, days, etc.). You can't collect some yearly data and some monthly data; all the "moments" must have the same unit of time.

As with any *longitudinal* case study, you must clearly specify the start and end points in the time series. However, because time-series analysis relies on *statistical* procedures and techniques, the definition of the time frame has added importance. In cross-sectional studies, including or excluding certain cases can introduce errors ("bias") that may reduce the validity of inferences or conclusions. The same is true, of course, if data for some of the moments (specific years, months, etc.) are missing.

One type of time-series analysis is *intervention analysis*, in which researchers want to see whether the values for a given variable change after a specific "intervention" (the independent variable). Because of the issue of *regression to the mean*, taking a snapshot of the year before and the year after is problematic, since we wouldn't know whether either (or both) of those years were outliers. The simple solution to this is to take several measures *before* and several measures *after* the intervention. Such a research design would look like this:

$$MMMMMM * MMMMMM$$

where each M stands for an individual measurement and $*$ represents the intervention.[2] There's no exact number of before/after measurements to take, but a good rule of thumb is six. Too many measures can introduce variation from other factors; too few may not be enough to get an accurate average for either time period. As always, these choices are up to you—but they must be clearly explained and justified.

Qualitative and Quantitative Research Strategies

There's a great deal of unnecessary confusion about the difference between—and relative merits of—qualitative and quantitative research. For one thing, many people confuse quantitative and *statistical* research: while statistical research is quantitative by nature, not all quantitative analysis is statistical; additionally, it's possible to use statistical procedures for some kinds of qualitative data. It's also important to remember that neither qualitative nor quantitative analysis is "better" (or more "rigorous") than the other. Both types of data/analysis have their strengths and weaknesses, and each is appropriate for different kinds of research questions. Finally, it's also important to distinguish between quantitative/qualitative *methods* and quantitative/qualitative *data*.

The simplest way to think about their difference is that quantitative data is concerned with *quantities* (amounts) of things, while qualitative data is concerned with the *qualities* of things. Quantitative data is recorded in numerical form; qualitative data is recorded in more descriptive or holistic ways. For example, quantitative data about the weather might include daily temperature or rainfall measures, while qualitative data might instead describe the weather (sunny, cloudy, mild). But these qualitative observations can be converted into qualitative measures if we start to count up the number of days for each descriptive. Or we might combine and/or transform our *nominal* descriptions into an *ordinal* scale (see Chapter 3). But we can also move in the opposite direction. For example, you could take economic data for a country, but instead of analyzing statistical relationships between the variables, you might instead describe the country as "developed" or "underdeveloped." This is especially appropriate if you were interested in researching the relationship "level of economic development" and some inherently qualitative concept, such as "type of colonialism" in either a single-case or comparative study.

Thinking about qualitative and quantitative *methods* is similar: Quantitative methods use precise, statistical procedures that rely on the inherent properties of the numbers involved. But this means that qualitative data, if transformed, can also be analyzed quantitatively. Qualitative methods rely on interpretative analysis driven by the researcher's own careful reasoning.

Qualitative Methods

Discussions about qualitative methods often focus on the method of *collecting* qualitative data. These can take a variety of forms, but some common ones include historical narrative, direct observation,

[2] This is a variation on the basic research design of measure, observe independent variable, measure ($M * M$).

interviews, and ethnography. Because much of this handbook focuses on quantitative methods, the discussion below is limited to brief overviews of a few major qualitative methods and approaches. The following descriptions are very brief, and focus primarily on implications for research design. More detailed descriptions of these methods, and how to *do* them are found in other chapters.

Historical Narrative. Perhaps the simplest (but by no means easiest!) qualitative method involves the constructing of historical narratives. This can be done through painstakingly searching through primary sources, which involves significant archival research. Not surprisingly, historical narrative is one of the basic tools of historians. Outside of historians—who prefer using *primary sources* whenever possible—social scientists often rely on *secondary* sources (analysis of primary sources written by other historians) to develop historical narratives. Beyond simply providing the necessary context for case studies, the data collection involved in constructing historical narratives is essential for *process tracing* analysis used in comparative studies.

Whether using primary or secondary sources, working with historical data requires the same kind of attention as working with any other kind of empirical data. You should treat the historical evidence you gather the same way you would a large-N quantitative study. In a large-N study, you must be careful to select the appropriate cases or make sure that important cases are not dropped because of missing data in ways that would bias your results. Similarly, using historical evidence requires awareness of missing data and other sources of potential bias. Additionally, since qualitative data is inherently much more subjective, it's important to use a range of sources to "triangulate" your data as much as possible. You should *never* rely on only one source for your historical narrative. Besides, summarizing one source is not "research." Instead, read as wide a range of relevant sources as you can and *synthesize* that information into a narrative, using the theory and conceptual framework that guides your research.

The main strength of historical research is that it can extend to almost any location and period of time. You are not limited by your ability to travel and "be there" to do research—although actually working in archives and other locations obviously strengthens historical research. You can also be creative about what constitutes "history" and historical "texts." Historical research can involve analysis of artefacts, material culture (including pop culture), oral histories, and much more.

The main weakness of historical research is that it often must rely on existing sources, which may have biases and/or blind spots. For example, a historian studying colonial Latin America has volumes of written records to choose from. But most of these are Spanish accounts (and mostly male), with few accounts from indigenous peasants or African slaves. Even more modern periods can be problematic: dictatorships, uprisings, fires, or even climate can destroy records. Good historical research involves making a careful inventory of what is available and being aware of what is missing.

Direct Observation. Unlike historical research, which can be done "passively" from a distance, direct observation requires being "present" at both the site and moment of research interest. *You*—the researcher—directly observe events and then describe and analyze them. One way to think about direct observation is to think of it like a traditional survey, except that instead of simply asking respondents some questions and recording their answers, you instead observe and record their behaviors.

Of course, direct observation doesn't have to involve human subjects at all; you could use direct observation simply to gather information about material items or conditions. The important thing is that direct observation is not the same as "remembering anecdotes;" direct observation should be planned out, with specific data collection strategy and content categories mapped out.

A major strength of direct observation is that because there is no direct *interaction* between you and the subject(s), it's more likely that the behaviors are "natural." Observational research can be done in a more natural setting, since there's no need to recruit participants or disrupt their activity in order to ask them a series of questions. Similarly, because you don't have to interact directly with your subject(s), there's a reduced change of introducing *bias* into subject(s) behaviors. Another strength of direct observation is that you're free to study behaviors in real time (an advantage of a natural setting) and you can also record contextual information (since *where* the behaviors take place matter).

The main weakness of direct observation is that you (the researcher) must be present to make the observations. For example, to study the Arab Spring uprisings using direct observation, you would have to have been present *during* the Arab Spring protests. Using newspaper reports and/or other people's recollections of the events is *not* "direct observation" (but a form of historical analysis). Also, because direct observation requires you to be present, this also means that you are limited to only the slice of "reality" that you are able to see at any given time, meaning that you need to think carefully about issues of selection bias. Even if you're directly observing a protest, you're only seeing it from your vantage point (in place and time). Being consciously aware of that is important.

Interviews. A non-passive, *interactive* form of research is personal interviews. While this can include a traditional *survey instrument* (which is generally described as a *quantitative* research method), typically by "interviews" we mean the more in-depth kind of conversations that use *open-ended* questions and allow more interpretative analysis. Interviews allow you to ask people with first-hand experience about events or expert knowledge about topics for detailed information. Even if you're simply using interviews as a way to get background or contextual information to help you refine your research project, interviews can be very useful.

Because interviews are an *interactive* form of research, they require approval by an institutional review board (IRB). Any interviews that you plan to use as data—whether in coded form or as anecdotes (quotations)—must be covered by an IRB approval *prior* to conducting the research. Among the things the IRB approval process requires is a detailed explanation and justification of your interview process, including how you will select your subjects and the kind of questions you plan to ask them. In addition to explaining how you will *recruit* your interview subjects, you will also need to specify how you will secure their consent. You will also need to explain whether the subjects' identities will be anonymous or not, depending on the scope of the research.

However, if you plan to use interviews as a primary research method—that is, if a significant part of your research *data* will come from interviews—then it's important to think carefully about interviews in the same way you would for other kinds of data. Because interviews are more time intensive than surveys, you do fewer of them. This means thinking very carefully about *case selection*: you want to be sure your case selection reflects the *population* you plan to study. This also means spending time lining up and preparing for your interviews. Lengthy interviews need to be scheduled in advance, and finding "key" subjects to interview can take a lot of effort, time, and legwork. And there's a lot more to interviews than just sitting down and talking to people; interviews require a lot preparation.

The advantages and disadvantages of interviews go hand in hand. Because interviews are open-ended, you can explore topics more freely. But that also means they take longer, you can do fewer of them. It also means they generate a lot of data, which you then need to sort through before you can analyze it. For certain kinds of research, interviews may be indispensable. Interviewing former politicians or social movement leaders may be a good way to study something as complicated as Bolivia's October 2003 "gas war." But finding the relevant social actors—and then scheduling

interviews with them—may prove difficult. At the same time, the memories and perspectives of the actors may shift over time, which is something to consider.

Ethnography. Ethnographic approaches aim to develop a broad or holistic understanding of a culture (an "ethnos") and are most closely associated with the field of anthropology, although they are sometimes also used in other disciplines (most notably sociology, but also political science). This approach involves *original* collection, organization, and analysis by the researcher. Ethnography can include unstructured interviews, but it often includes additional data collection. Perhaps the most common method of collecting ethnographic data is *participant observation*. Unlike the more "passive" observational research, in participant observation the researcher is an active participant, immersing him/herself in the daily life of his/her subjects. This, of course, requires transparency and *consent*: the population being studied must know that you are researching them, and must agree to include you in the group as a participant observer. The purpose of participant observation is to allow the researcher the ability to develop an empathic understanding of the group, and to describe and analyze the group from the inside out.

As an *interactive* form of research, ethnographic participant observation also requires IRB approval. Like with interviews, the IRB approval process requires you to provide as detailed as possible a description of the procedures you will use in your ethnographic research, including how you will handle and secure the confidentiality of your sources and data.

As with all other types of research, ethnography requires careful attention to sources of bias. Because ethnographic methods often rely on direct observations, you are limited to what you see. And because participant observation requires that your subjects (or "informants," in ethnographic lingo) know that you are observing them, this may alter their behavior, whether in conscious or unconscious ways. Fortunately, there are more *indirect* ethnographic methods that can be used to confirm (or "validate") observations.

The advantages of ethnographic approaches are significant: it can challenge assumptions, reveal a subject's complexity, and provides important context. The major disadvantages of ethnographic approaches have to do with limitations to access. Because many forms of ethnographic approaches require *contemporary* data collection and analysis, many tools of ethnography aren't available for historical problems (without a time machine, you can't conduct participant observation in the colonial Andes). Likewise, places that are difficult to reach, or where you have limited access do language or other barriers, are closed to you for many kinds of *direct* ethnographic approaches.

Quantitative Methods

Most of this handbook focuses on quantitative methods, but it's useful to at last sketch out two basic quantitative strategies for collecting data: surveys and working with databases. Like with qualitative methods, we can distinguish them between passive and interactive.

Surveys. Like open-ended interviews, traditional surveys with *closed-ended* questions are an interactive research strategy. Doing a survey requires interacting with people in at least some minimal way (even if only very indirectly through an online survey instrument). The difference between surveys and interviews, of course, is that you limit the kind of responses respondents can give (answers are "closed-ended").

It's important to remember that surveys are a large-N, quantitative research strategy. Because responses are closed-ended, the *quality* of the responses are shallow, which means you need to rely on their *quantity*. Surveys are only valuable if they're large enough to make valid inferences, if the samples are appropriately representative, and if the response options are validly constructed. But

just as interviewing is more than just sitting down and talking to people, conducting surveys is more than just making a questionnaire. In fact, designing the *survey instrument* (the questionnaire) is a critical part of survey-based methods. Surveys, like interviews, require IRB approval—and most IRB offices require a copy of the survey instrument. Any research design that includes a survey must also carefully outline how respondents will be selected or recruited, how many are needed/expected, and more.

Databases. All quantitative research is based on the analysis of a dataset, whether one collected by the researcher him/herself (this includes survey data collected, then organized into a database) or one prepared by someone else (such as the databases put together by your instructors for this course, which themselves were gathered and curated from various other databases).

Finding data from existing databases is the quantitative research equivalent of archival work. Just as historians have to be careful to select appropriate, credible sources, so too should researcher using databases. Whenever possible, be sure you should seek out the best, more respected sources for data. For example, most of the country-level data gathered by your instructors for this course comes from the World Bank Development Indicators, a large depository of data on hundreds of indicators (variables) for more than 200 countries and territories going back decades. There's a large (and growing) number of publicly available datasets made available by NGOs and governmental agencies, including publicly available survey data (such as from LAPOP and the World Values Survey).

The table below lists the six types of research designs discussed above along three dimensions: qualitative/quantitative, passive/interactive, and whether it generally requires IRB approval or not.

Table 2-1 Types of Research Designs

	Qualitative or Quantitative	Passive or Interactive	Requires IRB approval
Historical Narrative	Qualitative	Passive	No
Direct Observation	Qualitative	Passive	No
Interviews	Qualitative	Interactive	Yes
Ethnography	Qualitative	Interactive	Yes
Surveys	Quantitative	Interactive	Yes
Databases	Quantitative	Passive	No

Combining Qualitative & Quantitative Approaches

Just as you shouldn't limit yourself to only one kind of research design, you shouldn't restrict yourself to only one research method. Mixing different methods adds value to any research project. For example, you could combine a large-N survey with a few select in-depth interviews to provide greater detail. You could also combine historical narrative with ethnography. There are a number of creative ways to combine research strategies in "mixed methods" research that combine two or more different research methodologies.

One important reason for doing mixed-methods research is that it strengthens your findings' *validity*. Essentially, using two or more different strategies is a form of *replication* using different techniques. If

were using the language of statistical research, confirming a relationship between your variables in different kinds of methods could be described as "robust to different specifications."

Another important reason to consider a mixed-method research design is pragmatism. Although in theory, the ideal model of scientific research suggests that research design comes first, followed by data collection and analysis, the reality is that the process of data collection sometimes forces us review or original research design. If you have multiple types of data collection included in your research design, you can drop one of them if the data is unavailable. Likewise, if you discover that a type of data you hadn't considered could be incorporated into your research project, you should consider using it and adding another component to your overall research design.

A research design should be appropriate to your research question, and should help you leverage the best possible data. But it should also be flexible enough to accommodate the realities of research. Knowing how to do different kinds of methods allows you to adjust if new data becomes available or if expected data is suddenly unavailable (archives may be closed, interview subjects may prove too difficult to track down or recruit, or observation sites are inaccessible).

A Note About "Fieldwork"

Notice that this chapter hasn't mentioned "fieldwork." This is because fieldwork is best thought of as a *location* of research, rather than a *type* of research. While fieldwork involves going to a place and doing research there, it says nothing about whether the research is qualitative or quantitative. Some types of research require fieldwork by nature. You can't do observational research from a library (unless you are doing a study of behaviors in libraries). Although historians do much of their research in libraries, often those libraries are specialty archives located in various corners of the world. Even researchers who work primarily with quantitative data often rely on fieldwork. Some data is simply not available online, and must instead be sought out. Basically, if you *go somewhere* to collect data, you are doing fieldwork.

Being willing—and able—to do fieldwork is an important part of any researcher's toolkit. And whether the research is primarily quantitative or qualitative, all fieldwork requires careful planning and attention to detail. Most importantly, good fieldwork requires building relationships with a broader community of scholars and collaborators. Then again, the whole scientific process relies on building and expanding scholarly networks.

3 Working with Data

Whenever we do science, we work with "data." It's important to remember that "data" does *not* only mean *quantitative* data. Really, data just means "evidence." Both economic statistics and open-ended interviews are "data" because both are information that is collected, measured, and reported. But working with data also requires being aware of how to handle different *kinds* of data. "Facts" don't transform themselves into "data"; moving from observation to data is an intentional act. So learning how to "work with" data involves knowing how to transform observed "facts" into the kind of framework that can be used for analysis (qualitative or quantitative), and the various issues that this presents.

Operationalization

Earlier, we briefly discussed operationalization—the transformation of **concepts** into **variables**. This is a two-step process that involves **conceptualization** (clearly defining the concept) and the process of choosing or choosing specific measures for the variable. This second step is usually referred to as *operationalization*. This process involves more than simply deciding *how* to measure a concept, but also what *type* of measure; both involve deciding the *rules* for assigning measures.

Even concepts that seem simple to measure are complicated. How do we measure something like "size of the economy"? If you look around, you'll notice that there are a number of different measures for this: gross domestic product (GDP), gross national income (GNI), and gross national product (GNP). All three try to measure the same thing, but do so by including/excluding different things. GDP includes products and services *produced* in a country, GNI is the total domestic and foreign wealth produced by a country's citizens, and GNP includes products and services *consumed* in a country. And this before we start distinguishing between "real," "nominal," PPP (purchasing power parity), and various others adjustments to these measures. This is because there is no such thing as "the economy"—it's merely a social construction. Remember to avoid the danger of **reification**.

Other concepts are *much* more complicated. For example, how do we operationalize "democracy"? From political science, we know that democracy is a type of **regime** (a form of government). But is should we think of democracy as a **discrete** or **continuous** variable. In other words, are countries simply "democratic" and "not-democratic" (discrete) or can we place countries on a scale from most to least democratic (continuous). This is more than just a philosophical question, because different *types* of variables need to be handled differently. The key difference is that for continuous variables, each observation can theoretically take on any value between two specified value. Although continuous variables are more *precise*, this precision has to be justified conceptually. It's possible that precession may simply be an artefact of operationalization.

Before using a measure always go back to the original concept and ask yourself: Does this measure make sense for this concept? Your research design should include a discussion of—and justification for—the way you operationalize your concepts, as well as a discussion of the types of measures you use.

Levels of Measurement

The distinction between discrete and continuous variables/measures also has to do with distinction between levels of measurement. There are four levels of measurement: nominal, ordinal, interval, and ratio. Nominal and ordinal variables are discrete; interval and ratio variables are continuous. Although each level of measure is equally "useful" in different contexts, we typically think of levels on a continuum from "least" to "most" precise: nominal variables are least precise; ratio measures are most precise. Finally, it's important to note that we can move *down* the level of measurement, but not *up*. If you have interval-level data, you can transform that into ordinal-level data, but not vice versa.

Nominal

The simplest way to measure a variable is to assign each observation to a unique category. For example, if we think that the concept "region" is important for understanding differences across countries, we might categorize each country by region (Latin America, Europe, Africa, etc.). Because these measures are based on *ascriptive* categories, these are sometimes called *categorical* measures or variables. It's important to remember that nominal measures must place all individuals or units into unique categories (each observation belongs to only one category), and these must have no order (there's no "smallest" to "largest"). Although nominal measures are described as a "lower" level of measurement, this is only because they cannot be analyzed using precise or sophisticated statistical tools. Nevertheless, many important concepts (e.g. race, gender, religion) are inherently nominal-level variables.

One very specific type of nominal variable is a **dichotomous** variable. These are variables that can *only* take two values. A common example is gender, which we typically divide into "male" and "female," despite growing evidence that gender is fluid and non-binary. But dichotomous variables are useful in many instances. For example, if we simply want to measure whether a country had a military coup during any given year, but weren't interested in *how many* coups a country had, we could simply use a dichotomous variable ("coup" and "no coup"). Dichotomous variables can also be useful if we're willing to abandon precision to see if there are major differences between some breakpoint. For example, we could transform interval economic data into a simple "rich" and "not-rich" categories. In statistical applications, these are often called **dummy variables**.

Ordinal

Like nominal-level measures, ordinal-level measures *discrete* because the distance between the categories isn't precisely specified. Think of the difference between small, medium, and large drinks. Although these are *ordered* ("medium" is bigger than "small," but smaller than "large") the *distance* between them isn't necessary equal.

It's important to remember that ordinal measures are placed on an *objective* scale. The differences between small, medium, and large are ordinal because placing them on the scale says nothing about the *normative* value of small or large. For example, if we think of the variable for democracy as having only two categories ("democracy" and "not democracy") that's a *nominal* variable, because we have no *objective* reason to believe that democracy is "better" (I hope you agree with me that democracy is "better" than its alternatives, but this is a *normative* or "philosophical" position, not an *empirical* one). But this can be tricky: Imagine that we use the Freedom House values to come up with three categories: "free," "partly free," and "not free." In that case we *can* think of the variable as ordinal because we have categories arranged on a scale of freedom.

Interval and Ratio

If the distances between measures are both established and equal, then we have either interval or ratio measures. Once we know that the distance between 1 and 2 is the same as the distance between 2 and 3, we are able to subdivide those distances (1.1, 1.2, 1.3, …). That allows us a level of *precision* that's not possible with either nominal or ordinal measures. But that kind of precision is only possible if the distance between the measures is truly "known," and not just an artefact. Just because a variable is given in numbers, doesn't mean it's an interval or ratio measure. For example, the Freedom House and Polity indexes use numbers to places regimes on a scale from "most" to "least" democratic. But those numbers aren't "real," they're the product of expert coders who simply assign (although with a clear set of criteria) values to individual countries. In reality, those measures are *ordinal*. In contrast, something like GDP *is* an interval-level variable, since the distance between dollars (or yen, or euros, etc.) is precisely known. To speak of $1.03 cents has meaning in relation to any other price.

The only substantive difference between interval- and ratio-level measures is that ratio measures have an **absolute zero**. Typically, we think of an absolute zero as a value below which there are no measures. A simple example is age. Whether measured in years, months, days, or smaller units, a person can't be some negative number of years old. However, interval variables can also include money, which *can* go below zero (that's called debt). The reason is because the intervals between the units isn't just precisely known, they have a broader meaning. Take for example temperature. If we use a Fahrenheit scale, we can precisely measure the distance between 50° and 100°. But is the second temperature "twice" as hot as the first? Not really. Because there's no "true zero" in the Fahrenheit scale (although there is in the Kelvin scale, which has an absolute zero; on that scale the difference between 283.15° and 310.928° is almost trivial).

The table below lists the four levels of measurement, based on their distinguishing characteristics.

Table 3-1 Levels of Measurement

	Characteristics			
Level of measurement	Classification	Order	Equal intervals	True zero point
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Data Transformation

Working with data means more than just accepting data as you found it. It also includes the ability to *transform* data into other forms—particularly from one level of measurement to another. Just keep in mind that you can always move variables *down* a level, but never *up*. This can be done rather easily, but you have to take care to justify this in your research design. Sometimes we transform data for reasons that are guided by *theory*; other times we transform data for *practical* reasons having to do with the kind of analysis we want to be able to do.

For example, the Human Development Index produced by the UN comes as a *ratio*-level measure. There's an absolute zero (a country can't have "negative" development) and a maximum of 1.00. But how precise are the differences between each measure, really? Keep in mind that the index is *constructed* by combining a handful of economic, health, and education indicators into a single number. This is all done through a series of mathematical formulas that "force" the final number into something between zero and 1. How certain are we that the what we think is precision in the final HDI number isn't merely an artifact of the way the index was constructed? If we're not sure, we could decide to move down to a lower level of measurement. In fact, the UN anticipates this, and lumps countries by HDI score into four *ordinal* categories: very high, high, medium, and low levels of development.

Data transformation can also involve *altering* a variable in some way. But it's important that the transformation be *systematic*. If you alter a variable, you must do so for *all* the measures of that variable, not just a selective few. The only exception is if you have specific measures that are missing or problematic (you know they're "wrong"). But in those exceptional cases you must have a clear, transparent, and theory-driven justification.

Two common ways to transform a variable are to convert it to **z-scores** (see Chapter 4) or to use a **log transformation**. Briefly, a z-score transformation uses information about the way the variable is distributed (the mean and standard deviation) to create a new measure for the variable. This is only used in some specific situations (and in some ways as a matter of preference), which we won't go into here.

Log transformations are more common and should be in everyone's basic toolkit. Some variables are highly skewed (see next chapter) in ways that make comparing cases almost meaningless. For example, if we compare countries by population, China, India, the US, Indonesia, and a few other countries are simply orders of magnitude larger than the vast number of countries (many with populations below a few thousand). As you'll see later (in Chapter 6), using raw population measures would invalidate many forms of analysis. But the variable can be transformed using a **logarithm** of the original value. Simply, a logarithm is the *exponent* needed, for a certain *base*, to produce the original number. For example, for the base 10 logarithm of 1,000 is 3 because $10^3=1{,}000$. Unless you have very specific reasons to use a specific log base, the most common ones are base 10 and the "natural log" (which uses an irrational number e as the base).

Fortunately, you can do these transformations easily in Excel. For base 10, simply use:

```
=LOG(number, [base])
```

where number is the variable you want to transform and the optional command base is the base you want to use; if you leave that option blank and just use =LOG(number) then Excel automatically uses base 10. For the natural log, use:

```
=LN(number)
```

Measurement Error

Whenever we move from *concept* to *variable*, we are *constructing* data from abstract ideas in some way. This leads to potential problems of error, which has consequences for the **validity** and **reliability** of our data. There are two basic types of error in measurement: systemic and random.

Systemic Error

Systemic error is extremely problematic, especially if you're unaware of it. Sometimes, however, we are aware of systemic errors in our data. For example, we may know that some variable over- or under-estimates the true value of something. A classic example is unemployment statistics. In many countries (such as in the US), unemployment is measured as the percent the *actively engaged* workforce that is unemployed. What this means is that those who are unemployed but are *not* looking for work aren't counted in the unemployment statistics. That means we know that *actual* unemployment (if we mean "people without jobs") is *always* higher than the unemployment statistic. But we don't know by how much (and the discrepancy might change over time). This matters because a drop in the unemployment number can be a result of more people finding jobs (good) *or* a result of people giving up and no longer looking for work (bad). How we interpret the rise/fall in unemployment rate depends on what kind of systemic error you think exists.

Random Error

Random errors are simply "mistakes" made in measuring a variable at any given time. This can be problematic—or not—depending on how we interpret the random error. If random errors are truly random, then in any large sample over-estimation of the measure for one observation should be balanced by a similar under-estimation of the measure for another observation. In large-N cross-sectional analysis, this might not be a major problem—*if the random errors are relatively small*. In time-series analysis, however, such errors are problematic, since they make it different to observe real changes over time (random error might hide actual trends). But even in large-N analysis, if the random errors are *too large*, they may end up making the measures essentially meaningless.

Measurement Validity

The problem of measurement error has important consequences for the **validity** of measures. We can distinguish between three types of validity: content validity, construct validity, and empirical validity.

Construct Validity

Construct validity deals with the question of whether the operationalized variable "matches" with the underlying concept. We can begin to think about **face validity**, which simply asks us to consider whether the measure passes the "smell test." For example, if we operationalized "democracy" using the UN's Human Development Index, this would fail face validity. Democracy is a *political* concept, not an *economic* one. Although empirically we know that democracies are more likely to be rich than poor, a high level of socioeconomic development is *not* a criterion for democracy (unlike free and fair elections, the rule of law, etc.).

Content Validity

Another issue with content validity is that the measure should cover all of the conceptual dimensions of the concept. For example, democracy is a multidimensional concept that includes a number of things. If we develop a measure that only looks at some of them, but not others, we aren't really measuring democracy at all. For example, using mainstream democratic theory, Tatu Vanhanen (1984) developed an index of democracy that combined the dimensions earlier identified by Robert Dahl (1971): competition and participation. He operationalized *competition* as the proportion of votes won by the largest party from 100 (if the major party won all the seats, competition was zero); he operationalized *participation* as the voter turnout in that election. Although parsimonious, the measure never caught on because it ignored another important dimension: civil rights and political liberties. There's no "perfect" measure of democracy, and numerous types of indexes have

proliferated. Even the two most commonly used, Freedom House and Polity, have their own problems. Freedom House isn't actually a measure of democracy at all, but rather a measure of civil rights and political liberties (which can be a *consequence* of democracy, and therefore a useful **proxy** measure). The development of new empirical measure of democracy continues, and will probably never end. Largely because there's intense disagreements about the *content* (or conceptual *definition*) of democracy.

Empirical Validity

Empirical validity deals with the question of whether the variable measure is empirically associated or correlated with other *known* (or established) variables. This is sometimes referred to as **predictive validity**. We can test a new measure with an established or known older measure to see if they give similar estimates. If they do, then we can be confident that the new measure has empirical validity. Another way to discover this is to see if the measure for the variable we are interested is related with a different variable in a way that theory predicts. For example, imagine that we developed a survey questionnaire that asked people to define themselves along some dimensions that we then treat as a measure for "socioeconomic class." We could test this measure by comparing it to income (assuming we asked that of our respondents as well), since there's a strong (conceptual) relationship between income and socioeconomic class.

Measurement Reliability

The issue of measurement reliability is somewhat simpler. Here, we merely mean whether or not the measure gives *consistent* measures. For example, a scale is "consistent" if it gives me similar measures every day (assuming I don't loss or gain any weight). Let's suppose (because of vanity) that I reset the scale so that it's always 10 pounds lower than the real value. In that case my scale would be *reliable*, even though the measures aren't *valid*.

When you are developing your own measures, you can use some simple techniques to check for reliability: test-retest check, inter-item reliability check, and inter-coder reliability check.

Test-Retest

The test-retest method for checking reliability is pretty straightforward: take a measure multiple times, and compare them to each other (such as with the t-test explained in Chapter 5). Assuming you use the same procedures or decision rules, or collect the same kind of data, you should get the same (or at least *statistically* similar) measures. If you do, you can be confident that your operational measure is reliable.

Inter-Item Reliability

If your variable is a composite of multiple items, then you can check to see whether the various items are related to each other. For example, you could compare the four different indicators used in the Human Development Index measure and see whether each set of component indicator pairs is correlated. If the items are strongly related, then you can be confident that your measure is reliable.

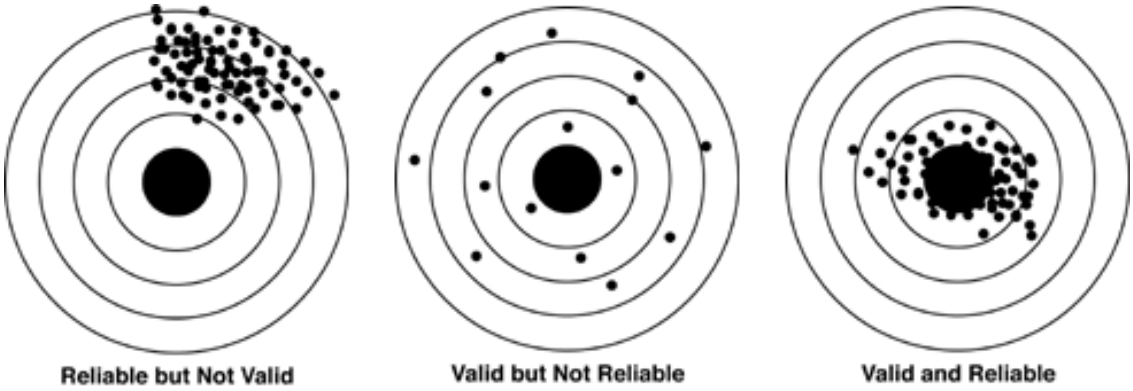
Inter-Coder Reliability

Finally, you can use other researchers (colleagues, assistants, etc.) to help check your measure's reliability by asking them to *independently* measure your variable. Then, you can check your measures to theirs. If you both get different measures, then something is clearly wrong: either one (or both) of you made an error or your measurement instrument is unreliable. This is a good test to use when you're working with a new type of measure that you're unfamiliar with. Even if you have no other

coders, you can simply "double-check" your measures yourself as a next-best option. The inter-coder reliability test is especially useful if your measures are a product of coding. For example, the Polity and Freedom House measure both rely on individual coders (experts on particular countries) coding the data based on some "coding rules" (often explained in a codebook). Ideally, these measures are first tested with small teams of experts who independently "code" the cases, assigning them the appropriate measures. If the coding rules are clear and understood by all the coders, they should all arrive at the same measures. If they don't, then the research team can review whether the error is a result of unclear coding rules, differences in judgement made by individual coders, or some other issue. A coded variable should only be used after it has successfully passed at least one inter-coder reliability test.

Measures are more reliable the smaller the errors (whether systemic or random). Although *validity* is in principle more important (since we want to be measuring what we think we're measuring), we can accept questionably valid measures if they are consistently *reliable*. That's because at least we can be confident that the relationships between variables we observe are "real" (since we can observe them across reliable measures). Over time, we may hope to learn how much error our measures have, and compensate for that. For example, imagine that you a shooting a rifle at a target. If you always miss, but your shots are clustered together, you have an inaccurate, but *reliable* rifle. Once you figure out how your shots group together, you can compensate and trust that, so long as you compensate for the systemic bias, you can hit the bullseye.

Figure 3-2 Validity and Reliability Compared



Source: "Validity and Reliability," Quantitative Method in the Social Sciences (QMSS) e-Lessons, Columbia University; http://ccnmtl.columbia.edu/projects/qmss/measurement/validity_and_reliability.html

Constructing Datasets

It's useful to think explicitly about how to actually *use* datasets. This is often overlooked in research training, and then new researchers make a number of silly mistakes and/or get frustrated trying to work with data. It's easy to think one only has to find and then download a dataset; but too often downloaded datasets are constructed in ways that aren't useful (after all, they were designed for a purpose other than the one *you* want to put them to). Beyond that, if collecting your own data (or even if *merging* data from various available datasets), you should have a basic idea of how to put together a dataset in a manageable form. Constructing a dataset in a systemic way will help you better keep track of your data and be able to use it. Lastly, the format I describe below is the one you'll need if you want to export your data from Excel into a statistical software package such as Stata or SPSS.

The first guideline is to distinguish between *variables* and *units of observation*. The conventional way that software packages handle data is to treat *rows* as observations and *columns* as variables, with the first row in a spreadsheet as the name of the variables. When you import any Excel spreadsheet into Stata or SPSS, for example, the software asks if you want to treat the first row as variable names. If you use that, then the software will use that text (or as close as it can) as the labels for the variables.

The second useful guideline is to make sure that the first column (on the far left) is for a variable that names each observation. Even if this isn't really a "variable" in the sense that you'll never use it for analysis, you should always try to keep the name (or unique code) of each observation as a running column on the far left. You'll notice that both the cross-sectional and time-series datasets have the names of countries running along the first column.

With the spreadsheet laid out this way, you're now ready to insert data. You can do this manually, or with copy and paste, just so long as you ensure that each row contains data from the same observation. On both the cross-sectional and time-series data for each cell in the *same row* is for data from the *same observation*.

A third useful guideline applies to the difference between time-series and cross-sectional datasets. For cross-sectional datasets, you can fit all the data in a single spreadsheet (each row a unit of observation or case; each column a different variable). For time-series data, however, you really have *three* dimensions in the dataset: unit of observation, variable of interest, and time. The simplest way to set up a time series dataset is to use a different spreadsheet for each variable (as you see in the class time-series dataset). In this case, each column would correspond to the units of *time*. A more complicated way (which is needed if you're going to use more advanced software for multivariate time-series analysis) involves treating the time-series data like cross-sectional data, but remembering that each unit of observation has *multiple* observations (so the cases are "country-year" rather than just "country").

If you have your data set up this way, you'll also be able to work with it in Excel to do all the various types of analysis described in the later chapters. You can always use blank sheets to run calculations, or even create new rows for items like means, standard deviations, etc. If you do that, however, it's useful to keep at least two blank rows between the last observation row and the row(s) for whatever descriptive or analytical statistics you plan to use.

A final note about datasets: It's good practice to start thinking about and constructing datasets early in the research stage. Too often, students spend a lot of time polishing their research design and literature review, before finally getting to the stage of collecting and/or organizing their data. This is a *big* mistake. Creating a dataset can take weeks or months (even years!) depending on the size and/or complexity of the data. New researchers can often end up caught in a quagmire unable to find and/or organize their data in a way that's useful for their analysis. When that happens, the analysis suffers in obvious ways that can't be hidden behind a flowery literature review.

4 Descriptive Statistics

If you use any kind of data, you need to present it in a meaningful way. Data (whether qualitative *or* quantitative) by itself is meaningless; it acquires meaning only through a conscious act by you (the researcher). One simple way to do that is through descriptive statistics, which summarize and describe the main features of your data. In any study involving quantitative data, it is a good idea to report or present that data in some way.

We often use descriptive or summary statistics, to *summarize* large chunks of data and present them in a meaningful way. Summary statistics typically report two types of statistics: measures of central tendency and of dispersion. These measures tell us something about the "shape" of the data. This information is then used to conduct *analysis*, which goes beyond merely describing the data to giving that data meaning.

Summary Statistics

One of the simplest ways is through the use of **summary statistics**. For example, an election in which millions of citizens voted, we obviously can't present a table listing the vote choice for each voter (since this would violate the secret ballot). We sometime can't even do that for smaller units (such as voting precincts). But even if we could, how useful or informative would that be? Including a complete, detailed dataset as an **appendix** might be useful, but it's not something that should be included in the main analysis. Instead, you should think about how to present a *summary* of that data that makes sense for your audience.

Below is an example of summary statistics for the 2014 Bolivian presidential election. Notice that is merely *summarizes* the national-level results for each presidential candidate by party. It also provides some information about valid, invalid, and blank votes, as well as the number of registered voters. But it also provides some percentages (or ratios) for those numbers.

Table 4-1 Votes by party in Bolivia's 2014 presidential election

Parties		*Candidates*	*Votes*	*Percent*
MAS	Moviento al Socialismo	Evo Morales	3,173,304	61.4
MSM	Movimiento Sin Miedo	Juan Del Granado	140,285	2.7
PDC	Partido Demócrata Cristiano	Tuto Quiroga	467,311	9.0
PVB	Partido Verde	Fernando Vargas	137,240	2.7
UD	Unidad Democrática	Samuel Doria Medina	1,253,288	24.2
Total Valid Vote			5,171,428	94.2
Invalid votes			208,061	3.8
Blank votes			108,187	2.0
Total votes			5,487,676	91.9
Registered voters			5,971,152	

Data from Órgano Electoral Plurinacional de Bolivia

Knowing the percent distribution of values in a sample or population is usually more useful than simply knowing the raw figures. For example, in 2014 more than one million Bolivians voted for Samuel Doria Medina, the candidate for Unidad Democrática (UD). But is that a little, or a lot? It might be tempting to simply compare it to the vote for the winner: Evo Morales, the candidate for the Movimiento al Socialismo (MAS), won nearly three times as many votes. But in another sense, we might also want to simply know whether the UD candidate dill well in comparison to other Bolivian elections or to candidates in other countries. If we did that we might notice that Doria Medina's 24.2% compares favorably to the 22.5% of Gonzalo Sánchez de Lozada, the 2002 candidate for the Movimiento Nacionalista Revolucionario (MNR), who won the presidency. It also compares favorably to the 20.6% of Lucio Gutierrez, who won the 2002 Ecuador elections. The fact that Doria Medina won over a million votes, or that this comes out to about a quarter of the total valid vote is simply a "fact" that has no *meaning* until it is placed into context. Summary statistics are a first step towards making sense of data.

One simple way to *transform* data in a way to give them meaning, is to use percentages (or shares). For example, we could transform the votes for Evo Morales into percentages simply by using a simple formula you should be very familiar with:

$$\text{Percent vote for party } X = \frac{\text{Vote for party } X}{\text{Total votes}} \times 100$$

Although you're probably used to thinking in **percentages**, many social scientists (especially when studying elections) prefer to use the term **shares**. The two numbers *mean* the same, but are slightly different. When you divide votes for party *X* by the total votes, you get the share of votes for party *X*. This number goes from zero to 1 (it won all the shares). To get a *percentage* as you're used to, simply multiply that number by 100. This may seem trivial, but it's important to remember the difference because if you treat shares as percentages, then the number 0.1 looks much smaller than it really is (10%). The best thing is to be consistent: either always use percentages, or always use shares.

Keep in mind that the *denominator* (the number at the "bottom" of the division) is very important. Evo Morales won 61.4% (or 0.614 share) of the *valid* vote in the 2014 election. This is the result reported by the the Órgano Electoral Plurinacional (OEP), Bolivia's electoral court. But you could also calculate this instead over the total votes cast (which would include blank and null votes), bringing Morales's vote share down to 0.578 (or 57.8%). And if we used the total registered voter population as the denominator, the vote share is 0.531 (or 53.1%). Which is still remarkably impressive: in 2014, more than half of all *registered voters* in Bolivia voted for Evo Morales.

But using percentages is also an important way to make useful comparisons across different cases. The differences in sizes (of the denominator) across countries often makes comparisons without using shares or percentages meaningless. For example, if we wanted to talk about "oil producing countries," who should be on the list? We could look at the countries that produce the most oil, and we would find that these are (in rank order): the US, Saudi Arabia, Russia, China, and Canada. In fact, by itself the US produces more than 15% of the world's oil. Other than Saudi Arabia (and maybe Russia), we probably don't consider the other countries as "oil producing countries." Part of the problem is that while the US and China are large oil producers, their economies are so large that the oil plays a relatively minor part in it. Why not control for size of economy by using oil rents (the money generated from oil production) as a percentage of GDP and then see which countries are the top "oil producing countries;" we would find that the new top five list now includes Congo, Kuwait, Libya, Equatorial Guinea, and Iraq. That list makes more sense.

Measures of Central Tendency

Measures of central tendency merely tell you where the "center" of the data for a variable lies. There are three basic measures of central tendency: mode, median, and mean (or "average"). These are all measures for datasets—that is, for describing or summarizing the center of data for multiple observations (whether across many cases, or for one case measured across time).

Mode

The mode is the simplest measure of central tendency. It's merely the value that appears most often. The mode can be used for any type of data (nominal, nominal, interval, or ratio), but it's most appropriate for nominal or ordinal data. Interval and ratio data are much more precise, and so unless the dataset is *very* large, the mode may be meaningless.

You can find the mode by simply looking through the data very carefully and identifying the value that appears most often. Or you can use the Excel function:

```
=MODE(number1,[number2],...)
```

in which you insert the array of cells for all the observations of the variable of interest between the parenthesis. When you do that, Excel will simply provide the most common number. Note, however, that Excel requires you to use *numbers* for estimating the mode. This means you will need to transform your nominal or ordinal variables into numerical codes. For example, you could transform *small*, *medium*, and *large* into *1*, *2*, and *3*. And you could also transform a nominal variable like race from *white*, *black*, *Hispanic*, *Asian*, and *Other* to *1*, *2*, *3*, *4*, and *5*. Keep in mind that the number transformation for nominal variables is *arbitrary*.

For example, if we wanted to look at the world's electoral systems, we see that there's a wide variety of them. We find the mode, and see that list-proportional is the most common electoral system.

Median

The median is a more nuanced measure of central tendency. Here, it's the measure that *exactly* at the middle of the data. This means that one half of the data will fall on one side of the median, and the other half of the data falls on the other side. Because the median assumes that the data has an order, the median is only appropriate for ordinal, interval, or ratio variables.

You could find the median by arranging all the observations from smallest to largest (or vice versa) and then looking for the middle number. If there's an even number of observations, the median is the midpoint between the two middle-most numbers. Or you can use the Excel function:

```
=MEDIAN(number1, [number2], ...)
```

in which you insert the array of cells for all the observations of the variable of interest between the parenthesis. For ordinal variables, the median will most likely be one of the original values—unless the two columns in which median rests are tied, in which case the median may be a fraction. For example, for the values 1, 1, 2, 2, 3, 3 the median is 2 (the middle of the distribution); for the values 1, 1, 2, 2, 3, 3, 4, 4 the median is 2.5 (midway between the categories "2" and "3").

If we look at the Human Development Index as an ordinal variable (with the four categories: very high, high, medium, and low), we see that the median is "3" (high). That means that half of the world's countries have "high" or better levels of human development, and half of the countries have

"high" or lower levels of human development. We can also compare this to the mode, which is also "3" (or "high").

Arithmetic Mean

Perhaps the most useful measure of central tendency is the **arithmetic mean**, sometimes referred to as the 'average." It is appropriate for interval and ratio variables; it is *inappropriate* for nominal and ordinal variables. Like the median, the arithmetic mean (or simply "mean")[3] describes the "center" of the data, but does so taking into account the full distribution of the data and the distances between each of the observational values.

The mean ($\bar{x}$) is calculated with formula:

$$\bar{x} = \frac{\sum x_i}{N}$$

where x_i is the value of each observation (the subscript i stands for "individual observation"); you sum up (Σ) all the observations, and divide by the total number of observations (N). You can also use the Excel function:

```
=AVERAGE(number1, [number2], ...)
```

in which you insert the array of cells for all the observations of the variable of interest between the parenthesis.

Let's look again at the Human Development Index, but this time treating it like a ratio variable (using the actual scores produced by the UNDP analysts). Applying the formula, we find that that the mean is 0.676. If we compare that to the median and mode, we find that the figures don't quite match up. The mean HDI score of 0.676 is about the HDI score for Egypt (0.678), which is in the "medium" category. Why don't mode, median, and mean match up? Remember that the mean is much more precise. But also because of the way the mean is calculated, it's highly influenced by outliers. As you'll see below, the information about outliers and how they relate to the mean also helps us calculate measures of dispersion (the "shape" of the data's distribution).

If you do not have the underlying data for a variable, but instead have the frequency distribution (or "aggregated" data), you can still calculate the mean. To do this, you simply have to take each value and multiply it by the number of observations (it's "weight"), using the formula:

$$\bar{x} = \frac{\sum fx}{N}$$

where f is the frequency of each value for x.

Imagine that we had frequency distribution data for the Fragile States Index along the 11-point scale, but not data for individual countries. We could use this to estimate the mean along the scale (for this example we'll assume the scale is interval, not ordinal). First, we multiply the frequency (f)

[3] There are three types of means: arithmetic mean, the geometric mean, and the harmonic mean. Most statistical applications simply use the arithmetic mean.

of each observation by its value (x), and then add all those values up and divide by the total number of observations (177 countries).

Table 4-2 Frequency distribution of Fragile State Index scores

Index value (x)	Frequency (f)	fx
11	4	44
10	10	100
9	23	207
8	38	304
7	33	231
6	21	126
5	12	60
4	13	52
3	10	30
2	11	22
1	2	2
N	177	1178

$$\frac{1178}{177} = 6.655$$

We can then check our estimated mean derived from aggregate data from the actual mean using disaggregated (individual observation) data, and we find that they're identical: 6.655.

Measures of Dispersion

While measures of central tendency help us understand the "average" value of a variable, they tell us little about the "shape" of the distribution. But we also want to know whether the values are highly concentrated, or widely dispersed. Three measures that help us understand the shape of the distribution are: standard deviation, coefficient of variation, and skewness.

These three measures of dispersion are all derived from the arithmetic mean ($\bar{x}$), however, which means they are only truly appropriate for interval and ratio variables. There are ways to describe the variation of nominal and ordinal level variables, but these are done qualitatively.

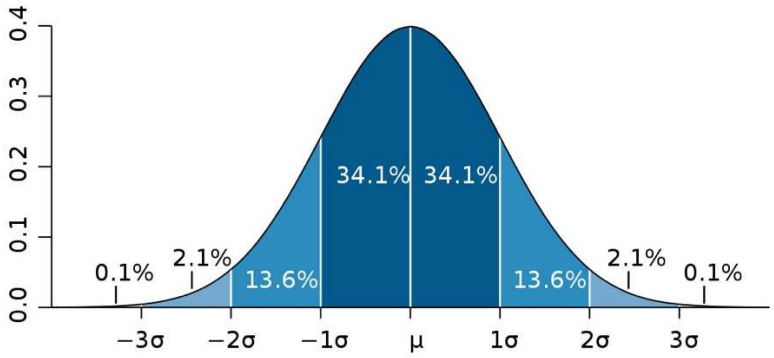
It's also important to note that these measures are best when the number of observations is at least somewhat large. Because the measures below use the arithmetic mean ($\bar{x}$) of interval level variables, they either assume a *normal* distribution or determine to what extent the distribution *deviates* from a normal distribution. In a perfectly symmetrical normal distribution, the mean, median, and mode would coincide. This is the "bell curve" distribution.

Standard Deviation

The simplest and most common measure of dispersion is the standard deviation. This measure assumes a normal distribution, and seeks to measure how widely the data is dispersed around the mean. Another way of thinking about this is that the standard deviation tells us how concentrated the data is around the mean.

Standard deviation helps us understand this because it is an abstract mathematical property: by definition, 68.2% of all the data fits within one standard deviation ($\pm 1\sigma$) from the mean and 95.4% of the data fits within two standard deviations ($\pm 2\sigma$) from the mean. The figure below shows a normal distribution of data, with marks showing up to three standard deviations ($\pm 3\sigma$) from the mean.

Figure 4-1 The normal distribution



Source: Jeremy Kemp, "Standard Deviation Diagram." Retrieved from "Probability Distribution," *Wikipedia* (https://en.wikipedia.org/wiki/Probability_distribution). Creative Commons license BY 2.5 (https://creativecommons.org/licenses/by/2.5).

Measuring the standard deviation depends on whether you are measuring it for a *sample*, or for a *population* (all of the possible units of observation):

$$\sigma = \sqrt{\frac{\sum(\bar{x}-x_i)^2}{N}} \qquad \text{or} \qquad s = \sqrt{\frac{\sum(\bar{x}-x_i)^2}{N-1}}$$

We use the Greek letter σ (sigma) to represent the standard deviation of *population*, and we use a lower case *s* for the standard deviation of a *sample*. In both cases, we subtract the value of each individual observation (x_i) from the sample or population mean ($\bar{x}$) and square that value. Next, we sum up (Σ) all the values of those subtractions. Then, divide that value by either the total number of observations for a population or by the number of observations minus one ($N-1$) in the case of a sample. Finally, we take the square root of that value.

To do this in Excel is straightforward, simply using the following command:

=STDEV.P(number1,[number2],...) ← for total population

=STDEV.S(number1,[number2],...) ← for sample population

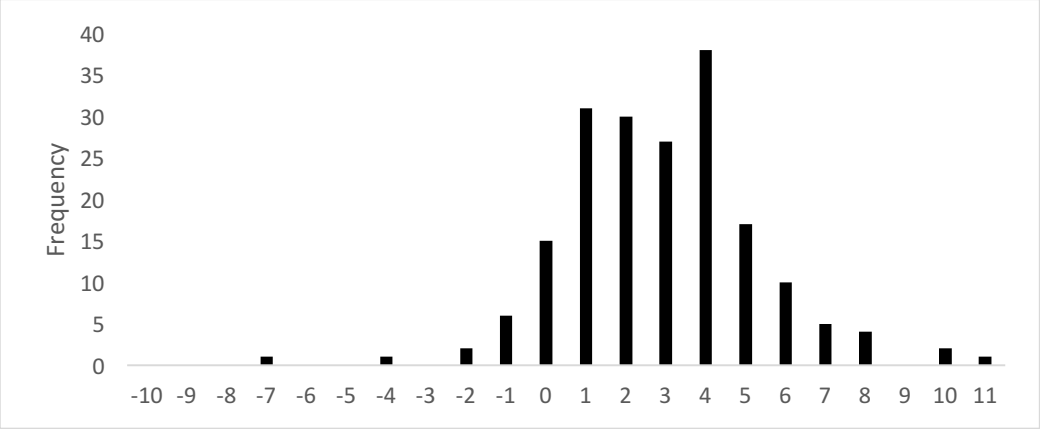
where number1,[number2],... refers to each individual observation. Or you can select a series of cells (an "array") in the same way as to calculate for the mean.

While the standard deviation is used in a number of other, more sophisticated forms of statistical analysis (often "under the hood"), it is useful for comparing *similar* observations. If you are comparing the standard deviation of infant mortality between two regions (Europe and Africa), the

differences in the size of the standard deviation help you understand whether the regions differ in how concentrated the

Let's look at the mean and standard deviation of GDP per capita growth from our dataset. Figure 4-2 is a histogram of the distribution of the variable GDP per capita growth across the 190 countries for which we have data. Notice that the numbers aren't perfectly distributed in a bell shape (like in Figure 4-1). But this is pretty close to a normal distribution, with most of the measures clustered around the mean (+2.38% GDP per capita growth).

Figure 4-2 Histogram of GDP growth per capita



We can also calculate the standard deviation for this variable, by simply using the Excel function and selecting the array for the observations. We find that one standard deviation is 2.41. If this were a *perfectly* normal distribution, we should expect that *exactly* 68.2% of the observations should fall between ±1 standard deviation (s) from the mean. So we should expect roughly that number of observations to fall between +4.79 (2.38+2.41) and -0.03 (2.38–2.41). When we check, we see that 138 (of 190) observations (or 77.4%) fall *between* those two extremes. Our observed data is a little different from an ideal normal distribution, but this is largely a product of the small sample size. In terms of statistical theory, 190 is a relatively *small* sample that can only *approximate* a normal distribution. Even if we study all of the world's countries (about 200, depending on how we count), we will rarely approximate a hypothetical normal distribution simply because our *population* is small.

Because interval/ratio data often resemble (or at least approximate) a normal distribution, one strategy for rescaling a variable is to use a **z-score**, which we can do if we know the mean and the standard deviation for a variable. All a z-score does is transform a variable so that *by definition* the mean becomes zero and the scale now runs ±1 unit for each standard deviation. A z-score for GDP per capita growth would make the mean zero and transform +4.79 into +1.0 and -0.03 into -0.03. The z-score is calculated with this formula:

$$z = \frac{x_i - \mu}{\sigma}$$

where μ is the mean (either sample or population, if known) and σ is the standard deviation (sample or population). You can do this automatically with Excel's STANDARDIZE function, which looks like this:

=STANDARDIZE(x, mean, standard_dev)

When you do this for a whole array of data, you'll notice that the mean is zero and the standard deviation is exactly 1.00.

Z-scores are often used to *standardize* different variables, which has application to many kinds of analysis. The advantage of a z-score is that the "units" for each variable are irrelevant (since we're just considering standard deviations). But the major disadvantage is that this makes *interpretation* of those results difficult, since you then have to go back and translate the standard deviation units back into the actual units for the variable.

Coefficient of Variation

A major limitation of the standard deviation, however, is that it is not useful for comparisons across different units, or even when two samples have very different means. For example, you can't compare the standard deviations of infant mortality and Human Development Index scores because the two variables have different scales. However, the coefficient of variation can only be used with ratio-level data for variables that have an absolute zero.

For comparisons between two very different variables (or if the means are very different), we can use the coefficient of variation, which is a *unitless* measure:

$$V = \frac{s}{\bar{x}}$$

The coefficient of variation is simply the standard deviation (of sample or population) over the arithmetic mean. While there's no function to do this in Excel directly, you can apply the formula in Excel like this:

```
=(stand_dev)/(mean)
```

by simply inserting the values directly, or selecting the cells that contain the values for the standard deviation and the mean.

Can only be used for ratio variables; can't take a negative number

Skewness

While standard deviation and coefficient of variation tell us about the "dispersion" of the values of a variable, there's a second element to the the "shape" of a variable's distribution around the mean. Skewness is a way of measuring where (and how much) the data for a variable "leans" in one direction or another.

Skewness can be calculated in many different ways. One of the most common—and the one used by Excel—is the following:

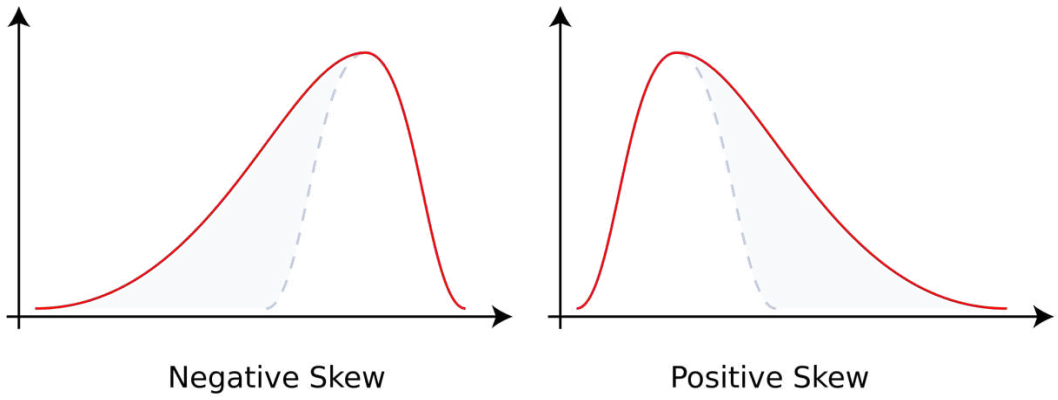
$$g_1 = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

To calculate skewness in Excel, simply use the following command:

```
=SKEW(number1,[number2],...)
```

where number1,[number2],... refers to each individual observation. Or you can select a series of cells (an "array") in the same way as to calculate for the mean.

Figure 4-1 Negative and positive skewness



Source: Rodolfo Hermans (Godot), "Diagram illustrating negative and positive skew." Retrieved from "Skewness," *Wikipedia* (https://en.wikipedia.org/wiki/Skewness). Creative Commons license BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)

Like the coefficient of variation, skewness is a *unitless* measure, which means you can compare the skewness of any two variables and compare them meaningfully. Unlike the coefficient of variation, however, skewness can be applied to any kind of ordered data (ordinal, interval, or ratio). Skewness is interpreted is as follows: If the data has a perfectly normal, symmetric distribution, then skewness is zero. A positive value shows that the data is *positively skewed*, which means that the tail is longer to the right of the mean. In other words, most of the observations are clustered at some point *below* the mean; the mean is higher than the median because a few outlier observations far to the right of the mean are driving the value up. Conversely, a negative value shows that the data is *negatively skewed*: the tail is longer to the left of the mean and most observations are clustered *above* the median.

When variables are extremely skewed, the standard deviation isn't very meaningful, which makes many kinds of tests of associations between variables difficult. One simple solution is to use the **log transformation** discussed earlier.

Reporting Descriptive Statistics

When reporting descriptive statistics, you should produce a table that lists the basic appropriate descriptive statistics for each variable. A common format for reporting is to report the mean, standard deviation, minimum, and maximum values.

Table 4-1 Descriptive statistics for selected economic sectors

Economic sectors as % of GDP	Mean	Standard deviation	Minimum	Maximum
Agriculture	13.2	12.36	0.0	55.4
Industry	29.3	13.41	6.6	77.2
Manufacturing	12.5	2.77	0.5	40.4
Taxes	17.1	7.74	0.0	55.7

Reporting the minimum and maximum values tells us something about the **range** of observations for the variable, which is a simple type of descriptive statistics. Because each of these variable use the

same unit (% of GDP), we can compare them. Notice that although agriculture and manufacturing have similar average values and ranges, their standard deviations are very different. Manufacturing seems to be more tightly concentrated around the mean.

To find the minimum and maximum values for each variable, you can simply rank order them and find the largest and smallest values. Or you can use the MIN and MAX Excel functions:

=MIN(array)

=MAX(array)

It's a good habit to always present your data into one (or few) descriptive statistics tables. You can also do this for qualitative data easily enough. There's no "right" way to organize a descriptive statistics table. It depends on the kind of data you are using, the type of analysis you plan to do, etc.

5 Hypothesis Testing

Many methods books refer to the following test statistics as "hypothesis tests," which is confusing because many other statistical procedures allow us to "test" hypotheses. But we begin with these because in some ways they're simpler. Basically, the test statistics presented here estimate ("test") the **probability** that an observed measure for *one variable* are the product of chance, rather than an actual relationship. They're also called **univariate inferential statistics**: they make *inferences* based on analysis of a single variable by statistically comparing two sets of data—or between one set of data and some hypothetical, known, or "ideal" reality—to determine whether those differences are meaningful.

There are two based types of univariate hypothesis tests: parametric and non-parametric tests. Most of these are *much* easier to simply "do" in a statistical software package (such as Stata, SPSS, SAS, or R). This handbook doesn't assume you have access to any of those, so it walks you through how to do them with Microsoft Excel. I've found that teaching this way forces students to wrestle with the underlying logic that makes these tests meaningful, and often gives them a better appreciation for *how* and *why* to use them.

Parametric Tests

Parametric tests are appropriate for *interval* or *ratio* variables, since it's easier to assume that they have *normal* (bell-shaped) distributions. If the variable measures are normally distributed (which we can test by estimating the skewness), then we can use a **difference-of-means test**, which uses the *mean* and *standard deviation* to compare between two populations, or between one population (such as a sample) and a hypothesized or known population (such as the "true" value). There are three basic kinds of difference of means tests, depending on whether you are testing one sample, two *independent* samples, or two *paired* samples.

All of these tests (as well as many other, more advanced statistical procedures) rely on estimating something called a ***t*-statistic**. It was developed in 1908 by William Sealy Gosset a chemist working at Guinness who needed a way to test the quality of the beer stout. Because company policy forbade him from making public trade secrets, he published his discovery under the pseudonym "Student," which is why the statistic is sometimes called a Student's t-test.

The *t*-statistic is a number that, by itself, is difficult to interpret. In the days before computers, you would have to calculate the value of t by hand and then look up a table that listed various values for t for different **critical values** and **degrees of freedom**.

Critical values are simply arbitrary percentage probability values set as the bar that must be cleared for a test to be meaningful. This is also known as the level of **statistical significance**, the minimum probability accepted for a test statistic to be "meaningful." The *minimum* level for statistical significance is usually .05, which essentially means that we can be 95% confident that an observed difference between the means is not due to random chance (because .05 means there's a 5% probability it is due to chance). However, many researchers prefer a *higher* threshold, so we typically report three different levels of significance: .05, .01, and .001. These are often thought of as

the **p-values**, but this is somewhat inaccurate. With computers, we can no calculate the exact p-value of a test statistic (we don't need to use tables anymore). Once we have a p-value, we simply look to see whether it is *smaller* than an established critical value (which is called "alpha"). This is why we tend not to report the actual p-values, but rather simply report whether p is smaller than some critical value (e.g. $\mathrm{p} < .01$).

The degrees of freedom is a number that tells us how much "freedom" our data has. Formally, it's the number of independent piece of information upon which a measure is based. Most commonly, the value for degrees of freedom depends on the number of observations (n) and the number of variables. For a one-sample test, the degrees of freedom is:

$$\mathrm{df} = n - 1$$

where n is the number of observations. For two *independent* samples, the degrees of freedom is:

$$\mathrm{df} = (n_1 + n_2) - 2$$

where n_1 is the number of observations in the first sample and n_2 is the number of observations in the second sample. The degrees of freedom for two paired samples is the same as for one sample, but in this case n stands for the number of pairs (not total observations).

One-Sample Difference-of-Means Test

The one-sample difference-of-means test has two basic uses. Because this test compares a *sample* to a *population*, it's commonly used to test whether a sample is *representative*. For example, if you collected data for a survey, and you wanted to know whether sample was representative, you could check to see whether it "matched up" with the population on various indicators. Your sample might not have the exact same mean as the population value (or "population parameters"), but you could check to see whether this difference was *significantly* outside what we might allow. The second application is basically the same: if you wanted to draw a smaller sample from some larger group, you could then test to see whether that group was significantly different from the larger sample.

The one-sample difference-of-means t-test follows the formula:

$$t = \frac{\overline{\mathrm{X}} - \mu}{s/\sqrt{n}}$$

where $\overline{\mathrm{X}}$ is the *sample* mean (the average value for all x's), μ is the known (or assumed) *population* mean, s is the standard deviation for the sample, and n is the total number of observations in the sample. However, if you know the *population* standard deviation (σ), you would be computing a z-test:

$$z = \frac{\overline{\mathrm{X}} - \mu}{\sigma/\sqrt{n}}$$

Since the components are easy to calculate, you could calculate this by hand and then look up the t value in a t-statistic table and use the information about the degrees of freedom and the described critical value to determine whether the sample was *statistically* different from the larger population. Or you can compute this with Excel and get the exact probability (or p) value.

The Excel Z.TEST function is used for all one-sample difference-of-means tests. For a one-sample difference-of-means tests in Excel you simply need to know the "true" population value, in addition to having data for a sample. If you also know the *population* standard deviation, you can also include this information. So if you know the population standard deviation (σ), then you're doing a proper z-test; if you don't know that information, then you're doing a one-sample t-test.

The Excel function for a one-sample difference-of-means tests looks like this:

```
=Z.TEST(array, x, [sigma])
```

where array represents the data cells for the sample, x represents the known population mean (μ), and sigma represents the *population* standard deviation (σ), if known. If the population standard deviation is known, then this is a true z-test; if the population standard deviation isn't known, then you can omit this from the function and Excel will simply use the *sample* standard deviation instead (making this a t-test). When you hit [RETURN] on the keyboard, Excel will give you the value for p.

However, this is a **one-tailed** difference-of-means test, and whenever possible you should use a **two-tailed** difference of means test. Remember that difference-of-means tests use information about means and standard deviations, assuming bell-shaped *normal* distributions. The two ends of the bell-shape are called "tails." A one-tailed test looks to see what the probability is that the sample mean rests at *one* of those tails. The one-tailed Excel Z.TEST is appropriate *only* if you specifically want to test the probability that the sample mean is *greater* than the population mean. There are very specific situations when a one-tailed test is appropriate, but social scientists prefer two-tailed tests whenever possible. Two-tailed tests actually make it *harder* to find statistical significance, because it *simultaneously* tests the probability that the mean is higher *and* lower than the population mean. In other words, the .05 critical value under the bell curve is split in half (each tail has 0.025 available).

There's no simple way to do a two-tailed one-sample difference-of-means test in Excel. But there is a way to do it with this slightly more complicated formula:

```
=2 * MIN(Z.TEST(array, x, sigma), 1 - Z.TEST(array, x, sigma))
```

Imagine that we wanted to test to see whether the level of human development (HDI) for the 19 Spanish- and Portuguese-speaking Latin American countries is significantly different from the rest of the world. Using our World Bank indicators dataset, we first estimate the mean HDI (0.68) and the standard deviation (0.159). Next, we separate out our 19 Latin American countries. We could also estimate the mean HDI for the region (0.72) and notice that it is slightly higher than the global average. Is this difference significant? Using the Excel z-test function, we could simply find an empty cell, and type the function, inserting the appropriate values for the population mean (μ):

```
=2 * MIN(Z.TEST(array, 0.68, 0.159), 1- Z.TEST(array, 0.68, 0.159)
```

This produces the value 0.3313, which means there's a 33.13% probability that the difference between the two means is due to chance. For social scientists, this is too high—it's *well* above the .05 minimum threshold.

Let's see what difference it would make if we *omitted* the population standard deviation (or if we didn't know it). In this particular case, we would use:

= 2 * MIN(Z.TEST(array, 0.68), 1- Z.TEST(array, 0.68)

This produces the value of 0.0223, which *is* significant at the p<.05 level. Why? Well, the standard deviation for Latin American HDI scores is very low (s = 0.066) compared to the higher population standard deviation (σ = 0.159). If we substitute the Latin America regional standard deviation, then the two means (0.68 and 0.72) are farther apart relative to the smaller standard deviation.

Let's compare this to the EU member nations:

= Z.TEST(array, 0.68, 0.159)

This produces a value of 6.1915E-9, which is negative exponential notation for 3.09757×10^{-9}, which is very *small* number (0.00000000309575) and well below the thresholds for statistical significance. Based on this test, we would say that the EU members have HDI levels well above the global average, and that we are confident at the p<.001 level. So our two tests confirm that Latin America is "average" in terms of global human development levels, but EU countries are above average.

As the standard deviations for your sample and the population get closer, the difference between a z-test and a t-test disappears. You *can* use a simple t-test. But if you know the population standard deviation, then you *should* use the z-test. A z-test has more statistical "power" than a simple t-test, since it's more precise.

Two-Sample Difference-of-Means Tests

There's another category of t-tests that allows you to compare two samples. There are two basic types: tests for *paired* samples and tests for *independent* samples. The test for independent samples compares two different samples or groups to see whether they are different from each other along one variable. The test for paired samples is often used to compare two measures taken at different times for a sample of observations. The paired-samples test could also be used to compare two different *variables* for one sample—but *only* if the two variables are of identical scale.

The Excel T.TEST function is used for three different versions of the t-test, and looks like this:

=T.TEST(array1, array2, tails, type)

where array1 represents the data cells for the first sample (x_1) and array2 represents the data cells for the second sample (x_2), with tails specifying whether you want a one-tailed or two-tailed test and type representing one of these three t-tests:

1. paired samples
2. independent samples with *equal* variance
3. independent samples with *unequal* variance

To select one of the three t-tests, you simply replace type with the corresponding number.

Two Samples with Unequal Variance. Unless you know that the two sample means have equal variances, you should use the test that doesn't assume equal variance. It's safest to simply always use the test that doesn't assume equal variance.

There are several ways to calculate t, depending on whether the sample sizes are the same size, and whether they have equal variances. Below is the formula for a Welch's t-test, which makes no assumptions about either equal variances or sample sizes:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s_1^2 is the squared standard deviation for the first sample, n_1 is the number of observations in the first sample, and $\overline{X}_1$ is the mean of the first sample; s_2^2 is the squared standard deviation for the second sample, n_2 is the number of observations in the second sample, and $\overline{X}_2$ is the mean of the second sample.

Imagine we want to compare whether the means for HDI index scores for EU countries and Latin America are significantly different. You could do that directly in Excel, with no prior calculations—although you will need to separate out the two samples (the simplest way to do this is to put them in separate columns. You would then type the following Excel command:

```
=T.TEST(array1, array2, 2, 3)
```

which uses a two-tailed test (replace tails with 2) and selects unequal variances assumption (replace type with 3). When you do this you should get a p-value of 8.8900-E09 or (0.00000000889). This is well below the .001 critical value, so we accept that Latin America and the EU countries have different HDI regional means.

Paired Difference-of-Means Test. The t-test for paired samples is meant to be used to compare two different observations (or measures) of the same sample observed at two different points in time. The most obvious way to use is to as a form of "panel series" analysis in which you have a measures for a group taken before and after some "intervention." Basically, you would consider the means of variable for the group in the first point in time and test whether the mean was significantly different from the mean for that variable in the second point in time.

Another way to use this test is to compare the means of two different variables—but only if they are similar in *scale*. For example, you can compare differences between male and female life expectancy (since they're on the same scale), but not life expectancy and GDP per capita.

In either case, it's *very* important that the two groups are "paired." So whether you're comparing means of one variable at two points in time or two variables, you must ensure that each data point for each variable is matched or paired with the corresponding data point for the *same observation*.

First need to calculate the difference between each pair of observations

$$d_i = y_i - x_i$$

and then calculate the mean difference ($\bar{d}$), and the standard deviation of the differences (s_d), which you will then insert into the following formula:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where n is the number of *pairs* (not total individual observations).

For example, imagine if you wanted to know whether, across Latin America, infant mortality was different between 1980 and 2010. Using the regional time-series dataset, we know that the mean infant mortality for our 19 countries in 1980 was 56.6 per 1,000 live births, which is much higher than the 17.5 per 1,000 live births. However, we also notice that the standard deviation for infant mortality in 1980 was 25.89, and in 2010 it was 7.76.

Using the Excel formula, we would type the following command:

=T.TEST(array1, array2, 2, 1)

which uses a two-tailed test (replace tails with 2) and selects paired values (replace type with 1). When you do this, you should get a p-value of 9.6304-E08 or (0.000000096304). This is well below the .001 critical value, so it's clear that infant mortality dropped across the region during the three decades since 1980.

Imagine we want to compare male and female life expectancy for the world's countries. Looking at the global cross-sectional dataset, we notice that male life expectancy is 67.2 years, compared to 71.9 years for women. Is this difference statistically significant? Using the Excel formula, we get a p-value of exactly 0.0000, below the .001 critical value.

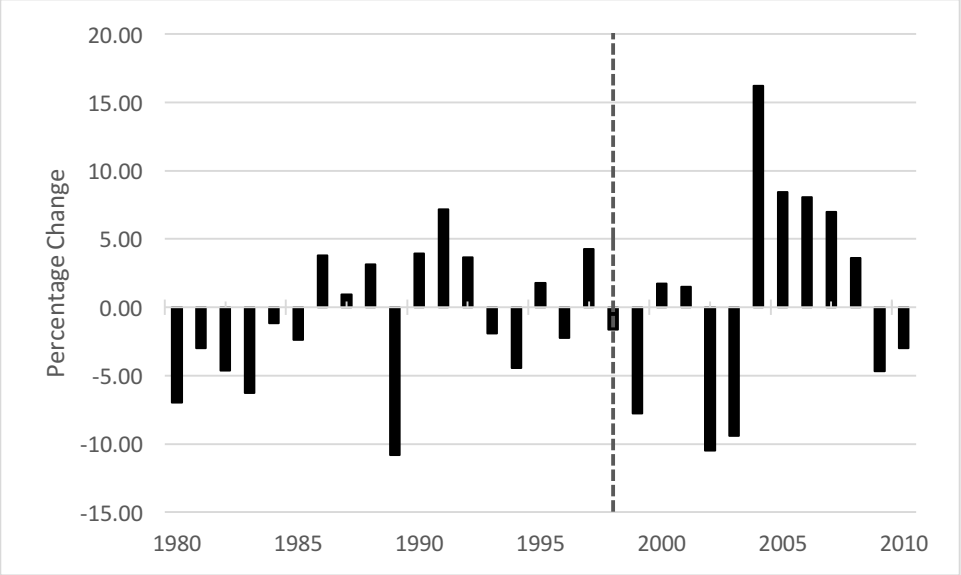
Using Difference-of-Means for Time-Series. You can also use difference-of-means tests for simple kind time series analysis. Because the family of t-tests can work for small samples, you can compare a relatively small number of observations before and after some event. Remember that the basic logic of time-series analysis looks like this:

$$MMMMMM * MMMMMM$$

where M is each observation in time and $*$ is some break in the time series; you can use any reasonable number of observations for each end of the time-series, but a good rule of thumb is at least six on each end. All you do then, is divide the time series around some "intervention" (either some specific event that happened, or even just a midpoint between two significant periods). Treating each half of the time-series as a different sample, you simply compare the means for the first and second periods.

For example, imagine we wanted to see whether Venezuela's economy improved after the election of Hugo Chávez in 1998. We could look at time-series data of Venezuela's GDP per capita growth. We notice that there's a lot of volatility across time, with many years of negative GDP growth, and some years of positive growth in the mid-2000s. If we use 1998 as a cutoff, we could look at GDP per capita growth between the periods 1980-1997 and 1999-2010. When we calculate the mean for each period, we find that the earlier period had an average growth rate of -0.84 percent, while the later (post-Chávez) period had an average growth rate of 0.94 percent. But because we know that means are sensitive to outliers, we want to know whether this difference is *statistically significant*. We can do this with a simple t-test for both periods.

Figure 5-1: GDP per capita (in constant 2005 US$) growth in Venezuela, 1980-2010.



When we do our two-tailed t-test we find that despite what looks like a large difference between the two means (average negative growth *vs.* average positive growth), the value for p is actually very high (0.5092). Basically, there's a little higher than 50% chance that the observed differences are a product of chance.

Reporting Test Results. All of the above difference-of-means tests are normally reported simply in the text where you discuss them. To report a t-test (or z-test), you need to report the t-statistic (or z-statistic), the degrees of freedom, and the level of significance.

Remember that the Excel functions we used above do *not* give you a t-statistic (or z-statistic) value, but the p-value. Fortunately, Excel has another function (T.INV.2T) that allows you to calculate the exact value for t. That function in Excel looks like:

=T.INV.2T(probability, deg_freedom)

To calculate t you need to know the degrees of freedom and the probability score for a *two-tailed* difference-of-means test (the p-value from the T.TEST function). You can calculate the degrees of freedom using the appropriate formula for calculating the degrees of freedom mentioned earlier.

Let's look at the last example (the time-series of Venezuela's GDP per capita growth). That was a t-test of two independent samples. The first sample was 1980-1997 (18 country-years) and the second sample was 1999-2010 (12 country-years). Using the formula for degrees of freedom for two independent samples we get:

$$\text{df} = (n_1 + n_2) - 2 = (18 + 12) - 2 = 30 - 2 = 28$$

If you plug the degrees of freedom value (28), as well as the value for p we obtained when we used the T.TEST function (0.5092) into the Excel T.INV.2T formula, you should get 0.668. So you should report the results of this t-test like this:

> There is no significant difference in Venezuela's GDP per capita growth in the years before the election of Hugo Chávez (1980-1997) and the years following his election (1999-2010); t (30) = .668, p=.509.

Because the results were *not* statistically significant, you should report the actual p-value. However, if the test *did* show a significant difference, then you should merely report the level of significance. In the earlier example of a *paired* difference-of-means test checking for differences in infant mortality across Latin America between 1980 and 2010, there *was* a statistically significant difference between the two samples. So you would report that like this:

> There was a significant difference in infant mortality rates across Latin America between 1980 and 2010; t (18) = 8.54, p<.001.

Non Parametric Tests

All of the above variations on the t-test are only relevant for variables measured at the interval or ratio level. If you want to do hypothesis testing for *nominal* (or "categorical") variables, you will need to use a non-parametric test. There are several different tests used in specific situations, which you can learn how to apply. This handbook will focus on one of the oldest and most common, which can apply in Excel: the Chi-squared test.

However, you should note that other tests that are considered more appropriate for different kind of nominal and ordinal data. These can be performed by most statistical software packages (SPPS, Stata, R, etc.). Because they are much more complicated to do "by hand" (and there's no simple way to do them in Excel), this handbook doesn't go over them in any detail. However, two of them deserve to be listed and briefly described:

- **Binomial test:** For dichotomous nominal variables, you can use an exact test of the *proportions* (the percent or share) of the two measures (e.g. 55% male, 45% female) between two populations. A simple application of binomial test would be to see if a coin is "fair" by comparing the number of times it comes up heads to the expected probability.
- **Ranked sum tests:** For ordinal variables, there's a variety of tests that can compare two samples (or one sample and the population) using the orders (the "ranks") of the measures to determine whether one sample tends to have larger values than the other. These are *inexact* tests, since ordinal variables don't have "true" means or standard deviations.

Although several other tests are either more common or more appropriate, you can use the simple Chi-squared test for many purposes. Remember: You can always go *down* a level of measurement. So you could do a univariate test of *ordinal* variables by transforming them into *nominal* variables (simply assuming there's no "order" to the categories) and then apply the Chi-squared test.

Once you understand the basic Chi-squared test, you will have a good understanding of hypothesis testing more generally, and shouldn't have any problem using the other tests.

Chi-squared Test

The Chi-squared (χ^2) test compares observed and expected values. Although it can be used like the z-tests and t-tests to compare one sample to a population or to compare two samples to each other, it can also be used to test *associations* between two nominal variables. For now, let's focus on using this test for *univariate* analysis.

The Chi-squared test uses the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed value for each cell and E_i is the "expected" value for those cells. For a simple univariate test, this is simple: the "expected" value is simply the known (or hypothesized) population or other sample distribution.

Let's walk through a simple example: Suppose you did a survey of 100 people, and you found that 60 of the respondents were female, and only two were male. You want to know whether this sample is "representative" of a population which in which gender is split 50/50. Because you will need to build a table in Excel for any kind of Chi-squared test, we can build one here for this simple example.

Table 5-1 Observed and expected distribution of male and female survey respondents

	Observed	Expected
Male	40	50
Female	60	50

To conduct our Chi-squared test, we would apply the formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \left(\frac{(40-50)^2}{50}\right) + \left(\frac{(60-50)^2}{50}\right) = \left(\frac{(-10)^2}{50}\right) + \left(\frac{(10)^2}{50}\right)$$

$$= \left(\frac{100}{50} + \frac{100}{50}\right) = (2+2) = 4$$

The value for χ^2 by itself isn't easy to interpret. Normally, you'd have to look it up on a χ^2 table to find the critical values for a sample of that size with that degree of freedom. Fortunately, the Excel function for a Chi-squared test (like the z-tests and t-tests) provides you with an exact p-value. The Excel function takes this form:

=CHISQ.TEST(actual_range, expected_range)

If you set up a small table in Excel that looks like the example in Table 5-1, you can easily select the correct ranges. For the example above, when you hit [RETURN] you should get a value for p of 0.046, which is just within the .05 critical value (but well over the .01 critical value); this sample falls within the 95% confidence interval for representativeness (but outside the 99% confidence interval).

When reporting the results of a Chi-squared tests, you are expected to report the the χ^2 value, the degrees of freedom, and the level of significance. If you report a table, you would include under the table (as a "note") the value for χ^2 and either the exact p-value or the range it falls under (in this case p<.05). However, if you aren't presenting a table, you would report the results of a Chi-squared test like this:

The sample is within the range for representativeness in terms of gender, χ^2 (1)=4.0, p<.05.

In this particular example, the degrees of freedom is one ($df = 1$), which is the minimum degrees of freedom we can have. Normally, however, the degrees of freedom for a Chi-squared tests is:

$$\mathrm{df} = (r-1)(c-1)$$

where r is the number of rows and c is the number of columns.

Let's look at an example in which the variable has more than two categories—and where it was originally an ordinal variable. Imagine we want to see whether human development levels in Latin America. We did this already with a t-test, using the numerical HDI scores. But we could also do this using the ordinal categories for human development used by the UN: very high, high, medium, and low levels of development. We may even have good reason to do this, since we could be skeptical of how *precise* the HDI scores actually are.

Using the named categories, we could construct a small table comparing the HDI levels for Latin America and the world:

Table 5-1 Human Development Index levels in Latin America and the world

	Latin America	World
Very High	3	48
High	10	53
Medium	6	41
Low	0	43

However, to use a Chi-square test to compare a sample to a population, we would need to compare the **proportions** (percentage shares) of both groups (Latin America and the world). When we do this, we get the following table:

Table 5-2 Human Development Index levels in Latin America and the world (proportions)

	Latin America	World
Very High	15.8	25.9
High	52.6	28.6
Medium	31.6	22.2
Low	0.0	23.2

Once we have our table, we can start to calculate the Chi-squared. We know the observed (Latin America values) and expected (world values). To apply the Chi-squared test formula:

$$\chi^2 = \left(\frac{(15.8-25.9)^2}{25.9}\right) + \left(\frac{(52.6-28.6)^2}{28.6}\right) + \left(\frac{(31.6-22.2)^2}{22.2}\right) + \left(\frac{(0-23.2)^2}{23.2}\right)$$

$$\chi^2 = \left(\frac{(-10.16)^2}{25.9}\right) + \left(\frac{(23.9)^2}{28.6}\right) + \left(\frac{(9.42)^2}{22.2}\right) + \left(\frac{(-23.2)^2}{23.2}\right)$$

$$\chi^2 = \left(\frac{103.15}{25.9}\right) + \left(\frac{575.18}{28.6}\right) + \left(\frac{88.68}{22.2}\right) + \left(\frac{540.25}{23.2}\right)$$

$$\chi^2 = 3.98 + 20.08 + 4.00 + 23.24 = 51.3$$

Then, using the Excel formula for the Chi-squared tests, we get 0.000 as the p-value. This is well below the .001 threshold, so we can say that Latin America is significantly different from the world. Whereas the world has a more "flat" distribution, Latin America has a more "normal" (or bell-shaped) distribution, clustered around "high" HDI level.

We can also use the degrees of freedom formula ($df = (4-1)(2-1) = (3)(1) = 3$) and use that to report our finding as:

> A Chi-squared goodness of fit test shows that Latin America is significantly different from the world in terms of human development level; χ^2 (3) = 51.30, p<.001.

Notice that this result confirms our earlier t-test. Also, note that when we use a Chi-squared test see if a sample is "representative" of a population, we are conducting a **goodness of fit test**. This and similar tests are reported in many more complicated statistical analyses. Later, we'll go over how to use the Chi-squared test for *bivariate* analysis. One final important note about Chi-squared goodness of fit test is that the *expected* distribution must include *at least* five expected frequencies in each cell.

6 Measures of Association

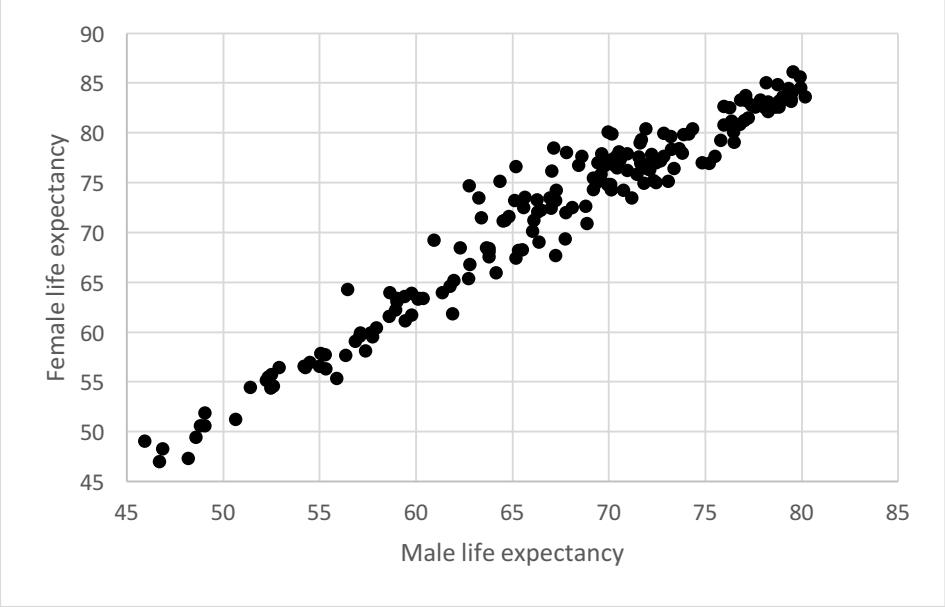
The following tests are typically referred to as **inferential statistics**, since go beyond describing variables to make *inferences* about the relationships *between* variables. Again, there are a large number of different kinds of statistical tools for analyzing various different kinds of relationships between two or more variables (and of different kinds of variables). If you understand the basic logic of inference, most of those techniques are fairly easy to understand. However, they require specialized software packages (Stata, SPSS, SAS, R, etc.). This handbook doesn't assume you have access to any of those, so it walks you through how to do some of them with Excel.

As with univariate hypothesis tests, the kind of inferential statistics analysis that is appropriate depends on the kind of variable you have.

Measures of Association for Interval Variables

With interval and ratio variables, we can use a wide range of statistical tools that rely on information about the the means and standard deviations. But perhaps the simplest way to understand a relationship between two interval-level variables is to plot them in a chart known as a **scatterplot**. This would simply plot each observation along two axes (x and y). Below is a scatterplot for the relationship between male and female life expectancy.

Figure 6-1 Male and female life expectancy scatterplot



The relationship looks pretty clear: in each country, male and female life expectancy are closely related. But how closely related? Notice that the data has a bit of a "bulge" as it goes up. So we know the relationship isn't very tidy. Fortunately, we can estimate the relationship more precisely with **linear regression**.

Linear Regression

Linear regression estimates the relationship between two interval or ratio variables. This is a simply algebra function that you probably remember as the one used to estimate the slope of a line:

$$y = \beta x + \alpha \qquad (\text{or } y = \mathrm{m}x + \mathrm{b})$$

where β is the **regression coefficient** (or "slope" of the line) and α is the y-intercept. Essentially, you're simply estimating that for every 1-unit increase in x what is the corresponding increase (or decrease) in y. In a scatterplot, you are trying to estimate the "best fit" line that goes through the scatter plot.

To estimate β you use the following formula:

$$\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Or you can simply apply the Excel SLOPE function:

```
=SLOPE(known_y's, known_x's)
```

Notice that for the SLOPE function you need to specify which variable is x and which is y. It usually doesn't really matter which is which, but be sure the slope formula and the scatterplot match: the x-axis is horizontal; the y-axis is vertical. Knowing which variable is x and which is y also matters for interpretation. When we apply this we find that the slope (β) for the relationship between male and female life expectancy is 1.10. Since male life expectancy is along the x-axis, we can say that for every additional year of life a man has, a woman in that same country should expect to live another 1.1 years.

Pearson's Product-Moment Correlation Coefficient

Linear regression only tells you the slope. But the slope might be the same regardless of how "tight" the data cluster along that same line. If we want to know how "strong" the observed relationship is, we want to estimate the **correlation coefficient**. The most common way to do this is with the Pearson product-moment correlation coefficient (also known as Pearson's r).

The Pearson correlation coefficient estimation uses the formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Or you can use Excel's PEARSON function:

```
=PEARSON(array1, array2)
```

Notice that in this case it doesn't matter which variable is x and which is y. This is because the Pearson correlation coefficient is only estimating the *strength* of the correlation between the two variables. The value of r can take on any value from −1 through +1. A negative value tells us that there is a *negative* or **inverse correlation** between the two variables (as the value of one variable increases, the value of the other decreases); a positive value tells us that there is a **positive**

correlation between the two variables (both values increase or decrease together). Although there's no "correct" way to interpret a Pearson's r value, typically we consider any value better than ±0.7 as a "strong" relationship. The strength of the relationship increases as the value approaches ±1.0. In our example for the relationship between male and female life expectancy, the value of r is 0.97, suggesting a *very* strong correlation.

Typically, we always report a p-value (or some other significance statistic) for any statistical test. Unfortunately, Excel doesn't have a simple function for the p-value of a Pearson correlation. To get the p-value, you'll first have to estimate t statistic for the Pearson correlation, using the formula:

$$t = \frac{r(\sqrt{n-2})}{\sqrt{1-r^2}}$$

where r is the value of the Pearson correlation coefficient and n is the number of observations. In Excel, that formula looks like this:

```
= (r(SQRT(n-2)) / (SQRT(1-r^2))
```

When we apply this to our example, we get the following:

$$t = \frac{r(\sqrt{n-2})}{\sqrt{1-r^2}} = \frac{0.97\left(\sqrt{187-2}\right)}{\sqrt{1-(0.97)^2}} = 54.2705$$

Once we have the value for t, we can estimate the probability value using Excel's T,DIST.2T function (which is a two-tailed tests):

```
=T,DIST.2T(x, deg_freedom)
```

When we try this, we get a value of 1.4213E-115, which is an incredibly small number. We can be *very* confident that there is a strong relationship between male and female life expectancy, and that this relationship is statistically significant.

You would report this finding like this:

> There is a strong positive correlation between male and female life expectancy; r = .97, p < .001.

Because the value of Pearson's r is always on the same dimension (from −1.0 to +1.0), you can easily compare any two correlations to see which one is stronger than the other. If we add information about the statistical significance, we can also make judgements about which relationships are more significant.

Linear Regression and Correlation with Log Transformation

Earlier we discussed how transforming variables sometimes facilitated analysis. One specific example was the use of **log transformation**, as a way to account for highly skewed variables. If variables are skewed (if the skewness measure is greater than ±1), they are candidates for log transformations.

For example, let's consider a possible relationship between doctors per 1,000 population and the child mortality rate (also per 1,000 population). We would expect to see a relationship between these two variables: all else being equal, fewer doctors should lead to more child deaths. A scatterplot of the two variables, however, looks odd: While it does seem like the two are related, the dots suggest a **parabolic relationship**.

Figure 6-2 Child mortality and doctors per 1,000 population

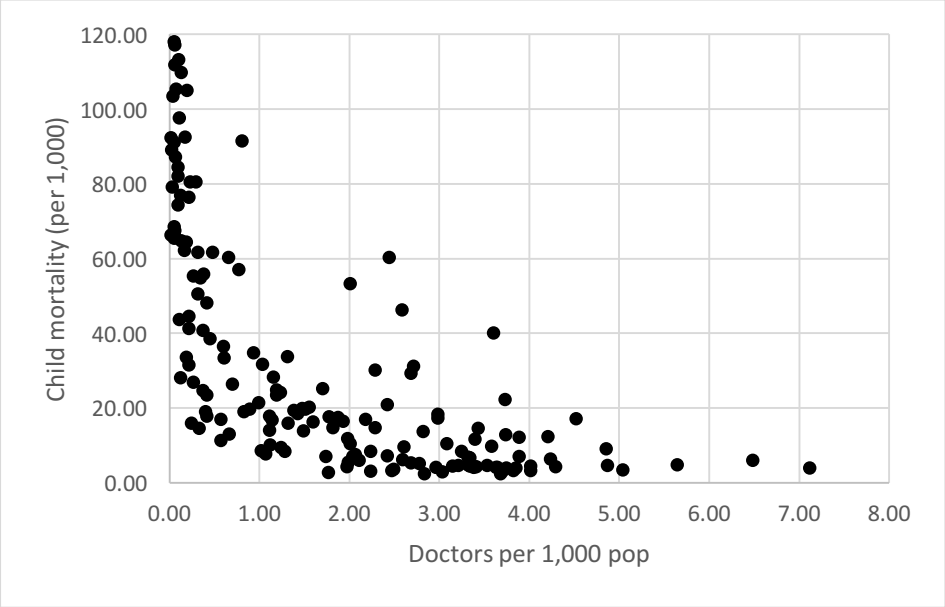
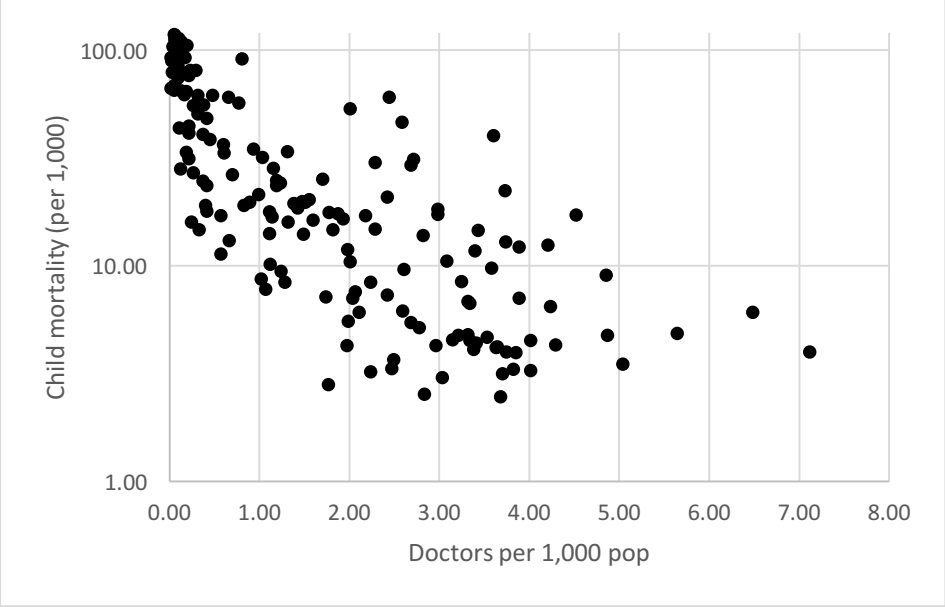


Figure 6-3 Log of child mortality and doctors per 1,000 population



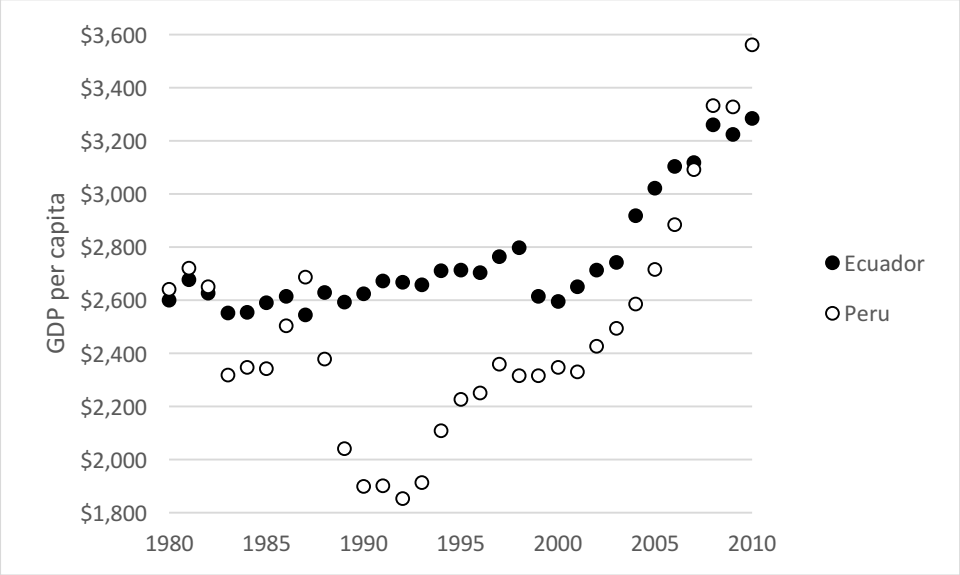
Compare Figure 6-2 and 6-3, which uses a base-10 log ($\log_{10}$) transformation for child mortality (which has a skewness of +1.33). Now the scatterplot looks a little more "normal" (although messy). If we estimate the regression coefficient of the $\log_{10}$ of child mortality and doctors per 1,000 population, we find arrive at r = -.77, which suggests a relatively strong inverse correlation. When we calculated for t we got a negative number (-15.73). Use the absolute value (15.73) and calculate the p-value, which is well within the $p < .001$ critical value. There is a mostly strong correlation, but it is statistically significant.

Linear Regression for Time-Series

You can also use simple regression for time-series analysis. Here, you would use a single variable of interest for y, and use time as the value for x. Otherwise, the procedure is completely the same, and all the comparative uses also apply.

Let's imagine that we want to see whether GDP per capita grew in Peru and Ecuador over time. Looking at the data tables, we see that both countries began 1980 with similar levels of GDP per capita ($2,600 for Ecuador and $2,641 for Peru). When we look at 2010, we see that the two countries again have similar values ($3,283 for Ecuador and $3,561). Just comparing those two numbers, we might think that Peru's economy slightly outperformed Ecuador's. But a scatterplot shows a slightly different, and more complicated story: Ecuador's economy seems to have stalled for about two decades (until about 2000), then grown rapidly. Peru's economy was volatile, but mostly falling, throughout the 1980s, then recovered and grew rapidly. Our scatterplot is very helpful for illustration, which allows us to make *qualitative* analysis of the two countries' economies.

Figure 6-4 GDP per capita in Ecuador and Peru, 1980-2010



If we connect the first and last data points, we notice that slopes for the two are different: Peru has a larger slope ($\beta = 23.25$) compared to Ecuador's ($\beta = 19.68$). But we can also use Pearson correlation to see how *strongly* time is correlated to changes in GDP per capita for each country. When we do, we find that the relationship between time and GDP per capita for Peru is modest and not statistically significant ($r = .50$, $p = 0.056$), but the relationship in Ecuador is fairly strong and statistically significant ($r = .81$, $p < .001$).

Looking at the scatterplot also suggests that we might want to consider breaking up the time-series into two different periods (1980-2000 and 2000-2010) because it looks like economic conditions improved in both countries since 2000. It also seems that the Ecuadorian and Peruvian economies performed very differently in the 1980-2000 period: Ecuador didn't see much growth, but at least it wasn't in freefall throughout the 1980s.

Partial Correlation

The examples above are all for **bivariate** correlation tests (they look at the relationship between only two variables). More common in social science are **multivariate** correlation tests. In these, we

estimate the effects of several independent variables on one dependent variable while simultaneously keeping each other variable constant. These are pretty straightforward in SPSS or Stata; if you understand the basics of regression analysis explained above, you can easily learn to use multivariate analysis techniques. In those analyses, reporting the correlation coefficient (β) for each variable is meaningful, and the software reports a p-value for each individual correlation coefficient, as well as an overall "goodness of fit" value, typically R-squared (which is just r^2).

However, there is one type of multivariate regression that is fairly easy to use. This is the **partial correlation**, which is a test that looks at *three* variables: a *dependent* variable, an *independent* variable, and a *control* variable. While this won't estimate a regression coefficient (β) for the independent variable, it does produce an easy to interpret Pearson's correlation coefficient.

The partial correlation uses the formula:

$$r_{YX_1X_2} = \frac{r_{YX_1} - r_{X_1X_2}r_{YX_2}}{\sqrt{1-(r_{X_1X_2})^2}\sqrt{1-(r_{YX_2})^2}}$$

where r_{YX_1} is the correlation coefficient for the relationship between the dependent and independent variables, r_{YX_2} is the correlation coefficient for the dependent and control variables, and $r_{X_1X_2}$ is the correlation coefficient of the relationship between the dependent and control variables. Basically, you need to first estimate *three* different correlation coefficients.

Let's go back to our example of male and female life expectancy. Let's suppose we want to control for gender parity in school enrollment as a way to control for gender social inequality. When we estimate all our correlation coefficients, we get the following values:

$r_{YX_1} = 0.97$

$r_{YX_2} = 0.56$

$r_{X_1X_2} = 0.52$

Once we have these values, we can plug them into the partial correlation formula:

$$r_{YX_1X_2} = \frac{0.97-(0.52)(0.56)}{\sqrt{1-(0.52)^2}\sqrt{1-(0.56)^2}} = \frac{0.97-0.2912}{\sqrt{(1-0.2704)}\sqrt{(1-0.3136)}} = \frac{0.6788}{\sqrt{0.7296}\sqrt{0.6864}} = \frac{0.6788}{(0.855)(0.831)}$$

$$r_{YX_1X_2} = \frac{0.6788}{0.7105} = 0.95$$

In the end, even when controlling for gender inequality, there's a strong relationship between male and female life expectancy. But notice that the value for r is *slightly* smaller when controlling for our gender inequality measure.

Measures of Association for Nominal Variables

We also have ways to test the association between nominal variables, all of which can be interpreted just like r. Some of these, however, require you to first calculate the Chi-squared statistic. Each of the measures of association differ, depending on the number of *categories* your two variables can take.

Phi Coefficient

If you have exactly *two* **dichotomous** variables, you can use the phi coefficient (ϕ), which is calculated with a simple formula:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Imagine we want to see if there's an association between electoral system and type of democratic government. But because we want to use the phi coefficient, we need dichotomous variables. So imagine if we only look at presidential and parliamentary systems, and compare that with list PR and first-past-the-post electoral systems. Our reduced sample of countries would look like this:

Table 6-1 Type of government and electoral system

	Presidential systems	Parliamentary systems
List-PR	29	33
FPTP	23	20

First, we would need to calculate the Chi-squared statistic. Unlike the earlier example, which was a sample test compared to a *known* population, here we are comparing the distribution to a *hypothetical* distribution—one that assumes there is *no relationship* between the two variables.

To estimate the *expected* distribution of the variables, we use the information in the *known* distribution to calculate the value of each hypothetical cell using the formula:

$$E_i = \frac{(\text{Total responses in row})(\text{Total responses in column})}{N}$$

This estimates values under the assumption that each row and each column has the same total observations, but that there's no relationship between the two variables (the assignment is, within the constraints of row/column totals, a 50/50 shot). When we do that, we would get the following:

Table 6-2 Expected distribution of type of government and electoral system

	Presidential systems	Parliamentary systems
List-PR	30.7	31.3
FPTP	21.3	21.7

Once we have this information, we can calculate the Chi-squared statistic using the formula we learned earlier:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

and we get $\chi^2 = 0.455$. When we plug this into the formula for the phi coefficient, we get:

$$\phi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{0.455}{105}} = \sqrt{0.00434} = 0.07$$

Remember that the phi coefficient is interpreted like a Pearson's correlation coefficient. So $\phi = .07$ demonstrates an incredibly weak relationship.

Lambda

If you have two nominal variables, and one of them can take on three or more values (categories), then you should use the **Guttman coefficient of predictability** (λ), which has the formula:

$$\lambda = \frac{\sum f_i - F_d}{N - F_d}$$

where f_i is the largest frequency within each level of the *independent* variable and F_d is the largest frequency of the totals for the *dependent variable*.

For example, let's say we expand our analysis of electoral systems and systems of government to include semi-presidential systems. We would see a distribution like this:

Table 6-3 Type of government and electoral system

	Presidential systems	Semi-presidential systems	Parliamentary systems	Totals
List-PR	29	12	33	74
FPTP	23	0	20	43

Lambda (λ) also requires you to specify which variable is the *independent* variable. Let's assume that we think government system is "dependent" on the type of electoral system a country has. So we would proceed like this:

$$\lambda = \frac{\sum f_i - F_d}{N - F_d} = \frac{(29 + 12 + 33) - 74}{117 - 74} = \frac{74 - 74}{43} = 0.00$$

Lambda is also interpreted like a Pearson's correlation coefficient, so a $\lambda = 0.00$ is a very weak relationship. It doesn't look like electoral system and type of government are associated.

Contingent Coefficient

In the event that you have two ordinal variables that have the *same number of possible* values, you would instead use the **contingency coefficient** (C), which uses the formula:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Again, you simply need to first create your observed table, estimate the hypothetical expected table, and use this to calculate the Chi-squared value. Then, insert that value into the formula. The contingency coefficient is also interpreted just like a Pearson correlation coefficient.

Cramer's V

If the two variables are "unbalanced" (one has fewer number of possible values than the other), then you need to use the formula to estimate **Cramer's V**:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

where k represents the smaller of the two values for each combination of variables (rows and columns in the distribution table). For example, if a table has 2 rows and 3 columns, then $k = 2$ (because $2 < 3$).

All of the nominal measures of association are reported in similar ways. You can either report them in the text with the basic format (just like for Pearson's correlations): describe the results of the test, then list the test statistic and its level of significance (from the Chi-squared test). For example:

> There is no noticeable relationships between form of government (presidential vs. parliamentary)t and type of electoral system (list PR vs. FPTP); ϕ = .07, p = .49.

Remember that for most of the examples for nominal variables, you would need to calculate the Chi-squared statistic, and for all of them you will need to find the significance level of the Chi-squared test statistic.

Measures of Association for Ordinal Variables

Things get even more complicated when start thinking about estimating the level of association between ordinal variables. The tests for nominal variables are inappropriate for ordinal variables because the order of the variables means that the *direction* of the relationship is meaningful. But because ordinal variables aren't as mathematically precise as interval or ratio variables, we can't use any of the tests for interval or ratio data.

These tests are not complicated, but they are cumbersome. There's no simple way to do these with Excel, so they have to be calculated by "brute force" (unless you use statistical software packages). But with a little bit of patience, you can estimate these easily enough.

Gamma

One test that we can use is **Goodman and Kruskal's gamma** (γ). Like the Pearson correlation coefficient, its values range from −1 to +1, which reflects the *strength* and *direction* of the association. The formula for Goodman and Kruskal's gamma (γ) is:

$$\gamma = \frac{Ns - Nd}{Ns + Nd}$$

where Ns is the number of "same-order pairs" that are consistent with a *positive* relationship, and Nd is the number of "different-order pairs" consistent with a *negative* relationship.

Imagine we want to test the relationship between levels of freedom and level of development across the world. We could arrange our observations for the Human Development Index and the Freedom House index, as in Table 6-5.

Table 6-4 Levels of freedom (Freedom House) and development (HDI) across 185 countries

	Low	Medium	High	Very high
Not free	11	8	6	3
Party free	29	18	21	5
Free	3	15	26	40

At a glance, it does look like there might be a relationship, but we need to make sure. The calculations for Ns and Nd aren't difficult, but they are a little tedious. To calculate Ns we start from the top left cell and look for all the "same-order" pairs; then do that for each cell, moving from left to right and top to bottom:

$$Ns = 11(18 + 15 + 21 + 26 + 5 + 40) + 29(15 + 26 + 40) + 8(21 + 26 + 5 + 40) + 18(26 + 40) + 6(5 + 40) + 21(40)$$

$$Ns = 11(125) + 29(81) + 8(92) + 18(66) + 21(40)$$

$$Ns = 1375 + 2349 + 736 + 1188 + 840 = 6488$$

$$Ns = 6488$$

Next, we calculate the value for Nd, which follows the same format, but in reverse:

$$Nd = 3(21 + 26 + 18 + 15 + 29 + 3) + 5(26 + 15 + 3) + 6(18 + 15 + 29 + 3) + 21(15 + 3) + 8(29 + 3) + 18(3)$$

$$Nd = 3(112) + 5(44) + 6(65) + 21(18) + 8(32) + 18(3)$$

$$Nd = 336 + 220 + 390 + 378 + 256 + 54$$

$$Nd = 1634$$

Once we have both Ns and Nd calculated, we can estimate gamma:

$$\gamma = \frac{Ns - Nd}{Ns + Nd} = \frac{6488 - 1634}{6488 + 1634} = \frac{4854}{8122} = 0.598$$

In the end, we discover that there is only a modest correlation between HDI level and Freedom House classification.

The one weakness of gamma is that it excludes any *tied* pairs. The more categories across both variables, the less likely there will be any ties. If there are only a few ties, then gamma can still be used, but it's accuracy decreases as the proportion of ties relative to the total sample increases.

If there are (many) ties, you can use a modification to gamma, known as **Kendall's tau-*b***, which is calculated with the formula:

$$\tau b = \frac{Ns - Nd}{\sqrt{(Ns + Nd + Ty)(Ns + Nd + Tx)}}$$

where Ty represents ties along the dependent variable and Tx represents ties along the independent variable. This is easier done with statistics software.

7 Advanced Inferential Statistics

The following is a brief description of some advanced inferential statistics that aren't easily handled with Excel; they require specialized statistical software. This discussion will focus on the abstract question of *when* these techniques should be used, and *how* they are carried out and reported. The discussion will also rely on discussions from Stata (the software I'm most familiar with) and SPSS (a software often available at university statistics labs).

This chapter explores four different techniques: multivariate regression, logistic regression, rank correlation, and binomial tests.

As with univariate hypothesis tests, the kind of inferential statistics analysis that is appropriate depends on the kind of variable you have.

Multivariate Regression

Perhaps the most common advanced statistical tests is multivariate regression, which is an extension of regression analysis to include two or more independent/control variables. And the most common version is known as **ordinary least squares** (or **OLS**) regression, which remains a "workhorse" technique in political science and sociology. Once you understand how OLS works and how it's reported, you should be able to quickly pick up more advanced forms of multivariate regression.

If you remember, the basic *bivariate* linear regression equation is:

$$y = \beta x + \alpha$$

In multivariate regression, we still estimating individual regression coefficients (β) for each individual variable. However, because there's now more than one independent variable, estimating each β also has to account for each of the other variables. To conduct a simple multivariate regression, simply select that test in either SPSS or Stata. The software will ask you to identify the independent and dependent variables. Remember, the dependent variable *must* use an interval or ratio level measure. But your independent variables can be any kind of measure: ration, interval, ordinal, or nominal (but only if the nominal variable is a *dichotomous* variable).

All multivariate tests produce a number of diagnostic indicators, many of which are rarely reported. In particular, SPSS output tends to generate a significant number of different statistics. The ones that are important and are generally reported are the following:

- Regression (or equivalent) coefficients
- Standard errors for each coefficient
- The level of significance (if any) for each variable
- Goodness of fit statistics
- The number (N) of observations

The output for multivariate analysis gives you a unique regression coefficient (β) for each individual independent (and control) variable. The SPSS output gives you both **standardized coefficients** and **unstandardized coefficients** (Stata allows you to select which one you want ahead of time). Standardized coefficients are based on standard z-scores for all the variables. This has the advantage of making it easy to compare the *size of the effects* for different variables on a universal sale (each 1 unit of change stands for one standard deviation). But this makes it difficult to provide a *practical* explanation of the effect of each variable using the variable's own scale (one unit of x leads to a one unit of y). I prefer to report *unstandardized* coefficients, but you can report either—as long as you're clear about which you use, and remember to interpret them correctly.

You should also report the **standard errors** for each variable. This doesn't apply to standardized coefficients, however, since z-scores make standard errors unnecessary. What the standard error tells you is the dispersion of each observation of x from the estimated slope line. The closer the standard error is to zero, the less likely the coefficient will be statistically significant. The standard errors are typically reported below the coefficients, in parenthesis.

Reporting the level of significance is typically done with little asterisk stars: one (*) for the $p < .05$ level, two (**) for the $p < .01$ level, and three (***) for the $p < .001$ level. These are recording next to the coefficients.

Finally, each model should report a **goodness-of-fit statistic** and the number of observations. The goodness-of-fit statistic for OLS linear regression is the **R-squared** statistic. This is a number that goes from zero to 1. The closer to 1, the better the goodness of fit. A simple way to interpret an R-squared statistic is to think of it as the share (or percentage) of the total variation in the *dependent* variable explained by the specific **model** (the combination of independent and control variables in the multivariate regression). By itself, an R-squared tells us nothing (any amount of explanation is better than not knowing). But we can compare the R-squared values of *different models* to see which one "performs" better. Generally, we prefer models that explain *more* with *fewer* variables (they're more *parsimonious*). And we always report the size of the sample (the "N").

Table 7-1 shows three different models, each considering the factors that affect per capita GDP:

Table 7-1 OLS correlates of GDP per capita (constant 2005 US$)

	Model 1	Model 2	Model 3
Industry as % of GDP	123.7		** 170.8
	(3093.05)		(52.85)
Labor force participation		-51.1	40.92
		(110.99)	(66.65)
Youth literacy rate			** 145.7
			(45.87)
Constant	* 7142.1	13414.8	* -15247.8
	(3093.05)	(7166.00)	(6744.31)
Number of observations	177	174	136
R-squared	.009	.001	.182

Unstandardized coefficients with standard errors in parenthesis; * p < .05, ** p < .01, *** p< .001

There are a few things to notice from Table 7-1: First, all of the necessary statistics are reported in the standard manner. Notice where the coefficients, standard errors, goodness-of-fit, and number of observations is reported. Also notice that the three models include a different mix of variables. OLS regression can be used for bivariate analysis (in which case it works like the examples in Chapter 6). The common use for multivariate regression is to create different different combinations of variables ("models"). This should be done guided by theory, however, and to develop an empirical argument. In Table 7-1 I tested industry alone (model 1) and labor force participation alone (model 2) to see if either of those variables had any significant correlation with GDP per capita. They didn't. But when I combined them along with a third variable—youth literacy rate—things changed: Now industry as % of GDP *was* significantly correlated with GDP per capita, as was youth literacy rate. The third model also had a much better R-squared value (those three variables alone explained nearly a fifth of the total variation in GDP per capita), while the first model had very weak R-squared values. So the weight of industry in the economy didn't seem to matter—except for when controlling for youth literacy (a proxy variable for level of education in society). Finally, the number of observations (N) in each model is different, because we can only regress the observations that have values for each variable; with no values, the observation is "dropped" (this is known as **listwise deletion**). It's also common to report the constant (the y-intercept for each model), although its statistical significance is not meaningful.

There are other advanced forms of linear regression, including ways to deal with time-series and panel data. Those are beyond the scope of this handbook. But once you understand the basic logic of the "workhorse" OLS regression, you should be able to learn the more advanced options easily enough.

Logistic Regression

If you remember, linear regression is only appropriate if the dependent variable is interval or ratio. But some variables of interest are nominal or ordinal. For example, if we might want to see what factors are likely to predict whether an individual votes, which is a binary variable (a person either votes, or doesn't), we need a tool to test for correlates of binary (or *dichotomous*) variable. For that, we use either **logistic regression** or the similar probit regression (both are very similar, but we'll limit discussion to logistic or logit regression).

It's important to note that logistic regression is *not* a form of regression on a variable that has been transformed into a log measure. The dependent variable *must* be a **binary nominal variable**.

Logistic (or "logit") regression is not strictly speaking a "linear" regression model. And instead of estimating a slope function, it estimates the *probability* function of a binary variable. Although logistic regression also produces coefficients for each independent/control variable, these aren't as easy to interpret as in the simpler OLS regression. For now, let's focus on simply knowing whether the coefficient is positive or negative (which tells us whether it increases or decreases the likelihood of observing the dependent variable) and whether the effect is statistically significant.

Logistic regression tables are reported much like OLS regression, with different columns for each model listing the coefficients, standard errors, levels of significance, and goodness-of-fit statistics. One major difference is that in addition to a "pseudo R-squared" statistic (estimated based on one of various procedures), you should report the Chi-squared goodness-of-fit statistic (usually reported significance level of "prob > Chi-squared").

Table 7-2 shows the results of three different models, each considering factors that predict whether a country is democratic:

Table 7-2 Logit estimates of probability that a country is democratic

	Model 1	Model 2	Model 3
Level of human development	*** 0.96		0.27
	(0.221)		(0.421)
Household consumption		*** 0.00	0.00
		(0.000)	(0.000)
Youth literacy rate			0.00
			(0.025)
Constant	*** –1.81	–0.53	–1.19
	(0.622)	(0.350)	(1.878)
Number of observations	120	110	79
Probability χ^2	0.000	0.000	0.010
Cox & Snell pseudo R-squared	.171	.246	0.133

Unstandardized coefficients with standard errors in parenthesis; * p < .05, ** p < .01, *** p< .001

Notice that the reported statistics in Table 7-2 are similar to those for traditional OLS regression. The one new addition is the Probability χ^2 reported as an additional goodness-of-fit measure. SPSS also provides two different pseudo R-squared estimates. You can use either one—but be sure to be consistent and to clearly label them.

Notice that among the independent variables are a mix of ordinal variables (HDI on a four-category scale) and two interval variables (household consumption and youth literacy rate). It may seem odd that household consumption was statistically significant with a coefficient of zero, but this may mean that the data is highly centered around the mean, making a small difference "decisive" in the difference for probabilities. It's also curious—and worth investigating—why the combined model has no significant predictors. But this is probably a result of having only 79 observations with data, which may introduce some systemic bias in the sample. It's worth testing this in various ways.

There are a number of advanced ways to use logit regression, not to mention its close cousin: probit regression. There's also a series of ways to use regression for ordinal variables, known as **ordered logistic regression** (and, of course, ordered probit). Those are also beyond the scope of this handbook. But once you understand the basic logic of logit/probit regression, you can explore those easily enough.

Rank Correlation

Earlier, when we looked at bivariate measures of association, we limited discussion to correlations between interval/ratio variables and nominal (categorical) variables. Here we focus on *bivariate* rank correlation tests (tests for a correlation between two ordinal variables).

These tests are known as rank-order correlation tests because they compare the paired rank orders of each variable for each observation. An ordinal variable that has three orders (e.g. small, medium, large); each observation ordered by the "rank" for each observation (e.g. 1, 2, 3). Since this repeats for the other ordinal variable, you can compare the "rank-order" of the two variables across each observation to see if there's a correlation between the rank orders.

One of the most common of these kind of tests is the **Spearman rank-order correlation** test. The correlation coefficient is known either as Spearman's rho (the Greek letter ρ or r_s), and is interpreted just like a Pearson's correlation coefficient (r): values range from ±1 (both variables are perfectly correlated) to zero (there's no relationship).

The formula for Spearman's rho is:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the two ranks for each observation. Like with Pearson's r, you can use r_s to calculate the value for t and obtain the statistical significance.

However, Spearman's rho can be used for interval or ratio data as well, which doesn't anticipate any ties. For ordinal data, you will have a lot of tie. That requires this other formula:

$$r_s = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

As you can see, this could be done with Excel—but for large datasets this can get very cumbersome. Fortunately, most statistical software (including SPSS and Stata) can easily handle Spearman's rho.

If we compare the four Human Development Index ordinal categories (1=low, 2=medium, 3=high, 4=very high) and the three Freedom House levels (1=not free, 2=partly free, 3=free) we get a value of 0.462. Always remember that even though these variables have numbers, the numbers are not meaningful (they are simply replacement for ordered categories): for example, a country with a HDI level of 2 ("medium") is not twice as developed as a country with an HDI level of 1 ("low)" or half as developed as a country with an HDI level of 3 ("high"). So, though you *could* estimate a Pearson's correlation coefficient (r) for these variables, you *shouldn't* because that test is only appropriate for interval- or ratio-level variables. Notice that this is consistent with our earlier test for this relationship using Goodman and Kruskal's gamma**.**

When you report a Spearman's rank-order correlation test, you report it just like you would a Pearson's correlation coefficient:

> There is a weak, but significant, positive correlation between human development and level of freedom; $r_s = .46$, $p < .001$.

More Advanced Statistics

There are many additional tests that are simply not covered in this handbook because they require specialized statistical software. But if you understand the basic logic of the various tests explained in this handbook, you shouldn't have any problem learning how to use them. There are many very good explanations of how to do many statistical tests in SPSS and Stata, which are the statistical packages available on most campuses. One very useful place for walk-through tutorials and brief, but clear and practical explanations is available from UCLA's Institute for Digital Research and Education (IDRE) available online at:

http://www.ats.ucla.edu/stat/AnnotatedOutput/

Another increasingly popular package is R. It has the advantage of being open source, but it has a relatively steep learning curve. Still, there's a growing number of books for beginning R users.

8 Content Analysis

Content analysis is a unique research method that merges qualitative and quantitative dimensions. Although it often relies on analyzing existing texts, it differs from "historical" research strategies that typically rely on narrative analysis. Content analysis transforms qualitative observations into *counted* observations. Content analysis can take many forms, both qualitative and quantitative. In the broadest sense, any type of analysis derived from communication—frequently written text, but also audio or visual communication (paintings or photography, film or audio recordings, etc.). In its simplest form, content analysis can take the form of consuming (reading, listening, viewing) some series of texts (newspapers, audio recordings, art exhibits) and presenting the interpreted meaning of those events to an audience. Those meanings are always "framed" by some sort of theory that gives shape and meaning to the content.

What Content Analysis Is ... and Is Not

It's important to distinguish "content analysis" (as a research method) from the traditional literature review process or the use of non-academic sources or texts (newspaper or magazine articles, films, performances, etc.) as reference citations in scholarly work. Content analysis involves a much more systematic process. While you are, in a very broad sense, "analyzing" the content of any reference materials in your research, you are typically doing so in a less intensive and more informal way. For example, when researchers use newspaper or magazine articles as additional references for key facts, figures, descriptions of events, or even statements by relevant subjects (politicians, social movement leaders, local residents, etc.), these are selected and many other similar newspaper or magazine articles are ignored. When doing content analysis, even newspaper or magazine articles that do not contain "citable" or "useful" information *are* analyzed, recorded, and included in the final research product.

It's also important to distinguish "content analysis" from traditional interview and survey research. While these are closer in structure to how content analysis is carried out, they're not as structured and systematic as most forms of content analysis. Another key element is that content analysis is usually reserved for "spontaneous" or "naturally occurring" communication—rather than the kind of solicited communication between an interview subject and researcher.

Content Analysis and Research Design

As with any method, there should always be a compelling and valid reason to use content analysis in your research, and this should be clearly stated. Prior to explaining the specific form your content analysis will take, you should provide a rationale for why content analysis is a valid way to answer your research question. This can range from the unavailability of other (perhaps preferred) data, to an argument that content analysis is "better" at addressing a specific research question and/or concepts than other methods, to using a different methodology to answer a question already posed by other researchers in a different way. You can also, of course, combine content analysis with other methodological techniques in your overall research design.

To be "social scientific," the specific technique used for content analysis needs to be clearly specified. This includes:

(1) Being explicit about the *theoretical framework* used and the *concepts* derived from that framework
(2) Being explicit about and justifying the *sampling frame* used to select materials
(3) Being explicit about the *unit of analysis*
(4) Being explicit about the way relevant concepts will be *operationalized* and measured.

Below is a descriptive sketch of a research design that uses content analysis to measure incidences of "coalition signaling" in Bolivian electoral politics through an analysis of newspaper reporting:

Table 8-1 Components of hypothetical research design

Theoretical framework and concepts	Theory: In parliamentary systems with many parties, parties campaign with an eye to future coalitions; they therefore send "signals" during the campaign process to potential coalition partners Concept: "coalition signaling"
Sampling frame	Newspaper reports of general election campaigns in major daily newspapers from 60 days prior to election through announcement of presidential election
Unit of analysis	Individual statement by each party's presidential candidate or party spokesperson(s)
Operationalization	Number of incidents when candidates or party spokesperson(s) did following: (1) acknowledged need for coalition to elect president; (2) mention rival candidates/parties, and whether this was positively or negatively; (3) mention ideological or programmatic similarities with rival parties; and (4) explicitly mention ideological or programmatic differences with rival parties

Harold Lasswell (1948) once described the basics of content analysis as determining "who says what, to whom, why, to what extent, and with what effect." In the above example, each article is read and coded in a particular way. The "who" for each statement is the "party" (whether a presidential candidate or other "official" spokesperson). The four variables measure or identify the "what" of the message. Theoretical assumptions guide the "why" and the "whom" of the message: the assumption is that even though statements by party candidates and spokespersons are probably primarily aimed at voters, statements about *other parties* or about *future coalition strategies* are intended to send "signals" (the "why") to other parties (the "whom"). The "to what extent" can be treated in two different ways, using **manifest** or **latent** analysis (see below); in this case, the statements could be analyzed in terms of the number of mentions ("manifest" analysis) and the strength (high/low) or direction (positive/negative) of their statement ("latent" analysis). Because the **sampling frame** included the final result (the naming of the president), the content analysis could also help answer the "to what effect?" dimension by allowing for a comparison between number, strength, and direction of statements about other parties and eventual coalition configuration.

Sampling Frames

Content analysis uses a similar kind of "sampling frame" research design as any other kind of large-N analysis. This is simply a more formal way of thinking about **case selection**—one shared with survey-based research. Before you can start to collect data on observations, you must first decide what is the "universe" of observations from which you will draw a **sample**. Your sample may include the whole universe of observations in your sampling frame, or a small subset of them.

For example, if you want to analyze how "the media" covered an election, you first need to develop a clear sampling frame—as well as a justification for using *that* frame. For example, "the media" is a broad concept that could include television, radio. Newspapers, internet social media (Facebook, Twitter, etc.) and more. Your sampling frame should be driven by *theory*, as well as *practicality*. Lack of access to radio and televisions transcripts or recordings of all the coverage (not to mention the sheer volume) may lead you to narrow your focus to newspapers. Even then, you will need to more narrowly defined your sampling frame: *Which* newspapers? During what *time period*? What *type* of coverage (front page, anywhere in the paper, exclude/include editorials, etc.). You should think through all of the potential questions, and explicitly walk your reader through your choices and your rationale for those choices.

Manifest Analysis

One simple way to do content analysis is to focus on *manifest* analysis. This involves looking at the objective (or "literal") meaning of the unit of communication under study. This often involves *quantitative* measures, such as counting numbers of stories, number of references to specific terms or individuals, or length of stories. We can then compare a series of observations (manifest analyses of different *units of analysis*) to others.

Even when manifest data does more than merely "count" events, references, or other markers—or employs other empirical or quantitative measures—it limits itself to the *obvious* meaning. Manifest analysis does not aim to provide interpretation of the "meaning" of the message itself. However, the difference between manifest and "latent" analysis (see below) can become blurred, particularly if we understand certain conventions of the medium as providing an additional layer of meaning.

Let's look at an example of the front page of *Página Siete* from Thursday, May 26, 2011 (Figure 8-1).

A first step towards manifest analysis could be to simply count the number of stories in the day's newspaper. If we include all "stories" found on the front page, we find 9 stories:

(1) Rising fares for *trufis* (the shared cabs used in La Paz)
(2) The Peruvian presidential runoff election
(3) The electoral law for judicial candidates
(4) Legalization (*nacionalización*) of illegal cars
(5) New ID cards
(6) Tornados in the US
(7) Controversy over TV *"cadena"* law
(8) Oruro mayor under investigation
(9) More cars with illegal license plates

Página
SIETE
Diario nacional independiente
Sin Tarjeta
iPhone 4
Sólo en Tigo, único proveedor autorizado en Bolivia
Trufis suben sus tarifas en Bs 0,50; la Alcaldía dice que es ilegal / 3
Vargas Llosa y 108 escritores se unen contra Keiko Fujimori / 11
Los medios no podrán entrevistar sin permiso del Tribunal Electoral
DATO Asambleístas del MAS y vocales electorales proponen formas para evitar el libre trabajo de los medios durante las elecciones judiciales. Pese al anuncio del presidente Morales, el MAS no reforma dos incisos de la ley electoral. / 2
Diputado Arce dice que las reformas son sólo para "mejorar la redacción".
El MAS desoye a ministro Arce y sigue legalización de autos chutos
CASO "El Ejecutivo tiene su posición y el Legislativo la suya, que puede no ser la misma, la norma es impulsada por el Senado". / 9
Plan de carnetización gratuita quiere ser base de la nueva Identificación
TEMA El responsable de la unidad que reformará el sistema, Antonio Costas, dijo antes que el sistema se basaría en el padrón biométrico. / 6
Más tornados afectan a EEUU
HURACÁN Al menos 14 muertos más en Oklahoma, Kansas y Arkansas. / 14
PRENSA / 2
Para la SIP es "abuso de poder" que la TV privada emita discursos de Evo
PROCESO / 16
Alcaldesa de Oruro activa su defensa política con ayuda de las mujeres.
AUTOS / 4-5
Aparecen en La Paz dos nuevos vehículos con placas irregulares.

Figure 8-1 Front page of *Página Siete* (May 26, 2011)

This level of analysis is very basic. But it allows us to compare this edition of *Página Siete* either with other day's editions (from the same paper), or with other publications, or a combination of both. Such a comparison would allow us to see if different publications cover different kinds of news or with different frequencies, as well as allowing us to track patterns in the kind of items covered (at least on front pages) of newspapers over an extended period of time.

Another element of *manifest* analysis that starts to add more complexity could include empirical measures of the *size* ("length") or *placement* of news stories. This somewhat blurs the line with *latent* analysis, but still limits itself to what is "literally" observed without making an effort to interpret the material.

For example, we could look up each of the nine stories listed on the front page and note the length—in words, paragraphs, "column inches" (a newspaper convention), or pages—given to each story. We could also note each story's placement (where in the newspaper it is located). Finally, we could also note whether the story was accompanied by any graphic elements (photographs, charts, etc.) or any other kinds of ancillary materials (for example, a "sidebar" with quotes or additional information). These elements help us make inferences about the significance of the story. But what distinguishes this from *latent* analysis is that the inferences are draw through a "filter" of pre-selected criteria that apply to any kind of story; these inferences are not drawn from any analysis of the *content* of the articles themselves. In fact, one can do empirically grounded and useful *manifest* analysis of material without even having to actually read the material at all.

Yet another way of doing manifest analysis is to look for specific references within a collection of materials, rather than analyzing the materials themselves. This does require reading of materials, but only for the purposes of looking for specific references. For example, we may want to look at a number of *Página Siete* (or other periodical) editions for references to specific people, words, or events.

Imagine we were looking for any references to President Morales or members of his government (the vice president and cabinet ministers or other important members of the administration). With manifest analysis, we would only count the number of mentions for each individual. We could count each story, or each individual mention. As with other forms of manifest analysis, we could also record the number length of stories that mention those figures, their location, or other readily observable features of the material in question. Such analysis could find, for example that certain cabinet ministers are mentioned more often, or that some are only mentioned in specific contexts (e.g. "National" news), while others are mentioned in a variety of contexts (e.g. "National" and "Local" news), or that some are mentioned alone but some are only mentioned with other individuals.

The type of manifest analysis used depends on the research question. Regardless, it is essential to clearly spell out in any research design or methodological discussion the *specific* parameters used to measure and report the findings of one's manifest analysis. This includes specific references not only to the kinds of material analyzed, but also the relevant time periods (for newspapers or magazines: what dates) that are part of the analysis.

Latent Analysis

A more complicated form of content analysis is *latent* analysis, which *does* require the researcher to use his or her judgment to infer *meaning* to material. This can range from a simple binary scale that rates stories as positive or negative, or a more complex form of analysis that looks about "quality" or

"depth" of the material. For newspaper material, a short story can have as much or more quality and/or depth as a longer story.

For example, we could look at coverage of one story (or "event") from several different newspapers and analyze the coverage along any number of dimensions. We can analyze whether specific "actors" (political figures, social movement leaders, etc.) are presented in a positive or negative light—or we could even go beyond a binary scale to create a more complicated *ordinal* scale along a positive-negative dimension. But we can also introduce other dimensions that we might think are important. For example, we could look at stories that deal with revolutionary change and determine whether the story (as a whole) and/or statements by actors cited in the story are framed in a "national-popular" or "Indian" tradition of rebellion.

The number and types of dimensions along which individual newspaper stories (or any other kind of material suitable for content analysis) are analyzed is unlimited. It's only important that a researcher states those dimensions clearly at the onset (in the discussion on methodology) and provides a clear **operationalization** of the kinds of phrases or other "indicators" used to place (or "score") any **unit of analysis** (whether a story, an actor's statement, or other pre-determined unit) along the stated dimension.

In addition to the above kinds of subjectively defined dimensions of analysis, we may also be interested in the quality of the article (or other communication) itself. For example, we may want to know whether one newspaper provides "better" coverage (of higher quality, with more contextual information, etc.) than another. This is essentially just another dimension, but here we are not interested in how the message is conveyed along some value dimension (positive-negative, democratic-authoritarian, local-national-international, etc.) but on a subjective evaluation of the *medium* itself.

What distinguishes latent analysis from the traditional uses of media (newspapers, radio, television, etc.) is in the *scope* of the analysis and how it is used. While traditional use of newspapers, for example, limits itself to the *selective use* of key articles used to provide evidence (often, anecdotal) in support of claims of fact or to bolster arguments, latent analysis follows the same conventions of manifest analysis: A sampling frame is determined, and *all* units of analysis included in the sample are subjected to the same kind of latent analysis, and that analysis is reported as a whole (only later are individual pieces selected for citations).

This means that, as with manifest analysis, a report using latent analysis should provide a table or other summary of the findings. This table would include the number of units of interest (e.g. individual articles, individual authors, or entire newspapers) analyzed, the dimensions used and the scores given to each unit are reported.

An Example: Analysis of Bolivian Textbooks

The following is an example of content analysis by a former student of the Bolivian field school program. In it, Leighton Wright analyzed Bolivian school textbooks to see whether their content had changed, reflecting the social and political changes following the election of Evo Morales. As part of her independent research project, Leighton analyzed a sample of available 4th and 7th grade social studies textbooks from time periods before and since Morales's election. Then, she developed a series of variables used to measure their differences across various dimensions (see Tables 8-1 and 8-2), including different indicators for "size" and topics covered. Using a fairly simple sampling

frame, Leighton was able to write an insightful analysis of differences in how textbooks represented Bolivia's ethnic diversity across several decades.

Table 8-1 Description of select 4th grade textbooks from 1989 to 2012

Editorial	Title	Year	Total pages	# of chapters	Civic ed. chapter	Lists each dept.	Represents indigenous peoples	Lists national holidays
Min. Ed. y Cultura	*Texto escolar integrado* (área urbana)	1989	98	21	No	Yes	Yes	No
Min. Ed. y Cultura	*Texto escolar integrado* (área rural)	1989	98	21	No	Yes	Yes	No
Don Bosco	*Ciencias Sociales Primaria 4*	2012	112	11	Yes	Yes	Yes	Yes
La Hoguera	*Ciencias Sociales Primaria 4*	2012	125	6	Yes	Yes	Yes	Yes

Source: Wright, Leighton. 2012. "The Effects of Political Reform on Identity Formation in Education."

Table 8-2 Quality of representation of indigenous peoples by textbook

				Quality of Representation		
Textbook	Grade	Year	# of pages	Low	Medium	High
Ciencias Sociales (Min. Ed. y Cultura)	4	1989	18	X		
Ciencias Sociales Primaria 4 (Don Bosco)	4	1989	21	X		
Ciencias Sociales Primaria 4 (La Hoguera)	4	2012	23	X		
El Mar Boliviano (Proinsa)	7	1988	0	X		
Lo positive en la historia de Bolivia (Proinsa)	7	1989	0	X		
Ciencias Sociales (Santilla)	7	1997	38		X	
Ciencias Sociales (*Lux*)	7	1998	6	X		
Ciencias Sociales (Bruño)	7	2012	10	X		
Ciencias Sociales (Don Bosco)	7	2012	78			X

Source: Wright, Leighton. 2012. "The Effects of Political Reform on Identity Formation in Education."

Leighton's study was a relatively simple one done with limited time (during the final week of a field study program), using "hard copy" (paper) materials. Certainly, given more time and using digital resources, she could've collected much more data and built a "large-N" dataset. If you use content analysis in this way, you can then use the data you produce in the same way you would use data from countries, surveys, or other data from any large number of observations.

Finally, there is advanced software for various kinds of content analysis. But simple content analysis tools are available to you already, if you have any kind of digital, "searchable" documents (PDFs, web pages, etc.): You can search a document to see how often terms appear in it. You can cut and paste text into Word and see how many words there are.

9 Specialized Metrics

So far we've focused on basic descriptive statistics (central tendency and dispersion measures) and inferential statistics (hypothesis testing and measures of association). But there's another category of measures that are useful, and which I refer to simply as **"metrics"** (ways of measuring). These can be very useful in the operationalization stage, as we move from concept to measure by transforming raw data into specialized indicators. Although there are a great number of these, I will focus on three: volatility, fractionalization (or "entropy"), and a special application of the fractionalization index used to measure the "effective" number of parties. If you have a sense of how these work, you can consider creative ways to use them in other contexts.

Even for the examples I provide below, there are a number of alternatives that are calculated in slightly different ways and produce different results. There are important methodological and substantive disagreements about which *specific* formulas are better and/or more appropriate to different contexts or purposes. Keep that in mind as you read the scholarly literature that uses such measures.

Volatility

Perhaps one of the simplest indexes is the volatility index, which measures the *aggregate* change in some variable across a range of cases from one time to the next. A similar term is used in financial economics, to measure the aggregate change in prices in a basket of stocks. In political science, a simple volatility index is often used to calculate the total *aggregate* change in votes across all parties, from one election to the next. This is called **electoral volatility**.

The electoral volatility index was developed by Mogens Pedersen (1979) as a way to measure the aggregate change in votes across elections for Western European democracies. Conceptually, Pedersen wanted to compare different countries along some dimension of party system "stability"; the volatility index allowed him to measure how stable voter preferences were between two elections for any country.

Electoral volatility is calculated as:

$$V = \frac{\sum|\Delta p_{i,t}|}{2}$$

where $\Delta p_{i,t}$ is the change in vote share for each individual party (i) at election t and the previous election t-1 (in other words: $p_{i,t} - p_{i,t-1}$). We take the absolute values of those subtractions, then sum them. We divide by 2 in order to avoid double-counting vote switches (our original step counts both the *added* and *lost* votes for parties). Basically, we're simply counting all the vote changes for each party to see how much voter preferences shift between one election to the next.

The advantage of the volatility index is that it is a standard "unit" of measure that can travel across any set of cases. Because V is calculated based on vote shares (fractions), the maximum value of V is

1 (100% of voters voted for a party other than the one they voted for in the previous election); the minimum value is zero (the vote shares between the two elections are identical).

For example, imagine a country with only three parties (A, B, and C) and their votes across elections were:

Table 9-1 Hypothetical vote share change

Party	Election 1	Election 2	Election 3
A	50	0	50
B	50	100	0
C	—	—	50
Volatility		0.50	1.00

In our hypothetical example, between election 1 and election 2, *half* of all voters (50%) "switched" from party A to party B, producing an electoral volatility of 0.5. Between elections 2 and 3, *all* voters (100%) switched away from B (to either A or C), producing an electoral volatility of 1.0.

If you have complete data for any pair of elections, you can easily calculate the electoral volatility with Excel. First, create a new column for each pair of elections in which you subtract one election from the other. The order doesn't matter, so long as you're consistent—but the convention is to subtract the earlier election from the most recent one. You can use Excel's ABS function to get the absolute value of each operation (each subtraction). Now you should have a column that matches up with each party, but only has the *difference* (the result of the subtractions) in the vote shares for each party. Note: be sure you include any party that only participated in one of the two elections (use zero for the election in which it was absent). Next, simply add up the values and divide by two (or multiply by 0.5).

As an example, we can calculate the electoral volatility between Bolivia's 2002 and 1997 elections:

Table 9-2 Change in vote share between 2002 and 1997 Bolivian elections

Party	2002 (t)	1997 (t-1)	Change (absolute value)
ADN	3.397	22.26	18.863
CONDEPA	0.372	17.16	16.788
LyJ	2.718	—	2.718
MAS/IU	20.940	3.71	17.230
MCPC	0.626	—	0.626
MIR	16.315	16.77	0.455
MIP/Eje	6.090	0.84	5.250
MNR	22.460	18.20	4.260
NFR	20.914	—	20.914
PS/VSB	0.654	1.39	0.736
UCS	5.514	16.11	10.596

Remember: we must include parties that didn't compete in one of the two elections (for example, MCPC ran in 2002, but not in 1997). We can also decide how to treat parties that change names

merge, or are "continuations" of other parties. For example, in the table above I treated MAS as a "successor" to IU (Izquierda Unida) because Evo Morales was elected as a congressional deputy representing IU, which was an alliance of several small leftist parties, including MAS. I did the same for Eje-Pachakuti and MIP.

First, we could calculate the change for each party (p_i). Next, to calculate volatility for the 2002 election (V_{2002}), we simply add up all the differences in vote shares, and divide by two:

$$V_{2002} = \frac{\sum(18.863+16.788+2.18+17.230+0.626+0.455+5.250+4.260+20.914+0.736+10.596)}{2}$$

$$V_{2002} = \frac{98.436}{2} = 49.218$$

We find that nearly half (49.2%) of voters "switched" parties between 1997 and 2002. By itself, this suggests a highly unstable party system. However, we can get a better sense of *how unstable* by comparing with other elections in Bolivia—as well as elections in other countries.

Note that above we calculated the *aggregate* national-level electoral volatility. It's also possible that electoral volatility at subnational levels (municipalities, single-member or "uninominal" districts, and departments) could vary significantly. These are areas worth exploring, and there's a growing literature in this area. You may also notice we've discussed volatility as a measure of changes in *vote shares*. But you can easily use this formula to measure differences in *seat shares* (the share of seats each party has in any election). Comparing seat and vote share volatility may also be informative about electoral politics in a country. Lastly, you can also use volatility to measures changes across other nominal variables (e.g. ethnic identification). The simple logic of the volatility formula is that it provides a simple metric that can be applied uniformly across cases and/or across disaggregated subunits of cases in a variety of ways.

Fractionalization

Another simple measure that can give a "number" to a dimension of data is **fractionalization**, which is a type of entropy index, a series of measure that look at the *inequality* of distribution of some variable. One of the most common entropy indexes is the **Gini coefficient**, which measures the level of economic inequality in a society.

One of the simplest measures of fractionalization is the **Herfindahl-Hirschman Index** (or **HHI**), which was originally developed in the 1940s as a way to measure marketplace concentration across a range of firms (i.e. how much the market for cars, for example, was concentrated on a few firms as opposed to dispersed among many). HHI is calculated as:

$$HHI = \sum s_i^2$$

where s_i is the share of each individual unit (which can be party, ethnic group, occupation category, etc.). As HHI approaches 1, the "market" is highly concentrated (a measure of 1 means that only one group exists); as HHI approaches zero, the "market" is highly fragmented (a measure of zero means that every individual in the sample is unique).

The simple HHI is based on "sum of squares" mathematics, which derive from the inherent properties that these have (if you recall, regression analysis uses squares). Recently, a number of

other indexes have been developed using the HHI as a building block. In particular, there are measures for ethnic fractionalization and the "effective" number of parties.

Ethnic Fractionalization

One application of this measures was developed by Alberto Alesina and several coauthors (2003) to measure the level of ethnic fractionalization:

$$F = 1 - \sum s_i^2$$

This formula simply transforms the HHI "concentration" index into a "fractionalization" index by subtracting HHI from 1 so that zero means a perfectly homogenous population (all individuals belong to the same ethnic group) and ethnic diversity increases as the number approaches 1 (a maximum value of 1 would mean that every individual belongs to a *different* ethnic group).

Because this measure offers a *universal* (and abstract) "unit" of measure, it can be used across any cases (or across subunits of a case) for informative comparison. It also means that a highly *qualitative* variable like "pluralism" or "ethnic diversity" can be given an interval measure, opening up the ability to use an otherwise nominal variable for a wide range of precise statistical analysis. In doing so, of course, it's important to remember to be careful for **reification**: the measure is *not* the concept; it's simply a mathematical artefact. Additionally, the indicator is only as good as the underlying data.

Finally, remember that just as Alesina took an indicator used in market economics and applied it to ethnic diversity, you certainly are free to use the fractionalization index to measure other nominal variables.

Effective Number of Parties

Another application of the fractionalization index is as a way to "count" the **"effective" number of parties** in a country. Most countries have a number of political parties. Even the US is not in this sense a "two-party" system (there are the Green, Libertarian, Socialist, and several other parties that most Americans never vote for). And in each country, some parties are "bigger" than others. A while ago, political scientists were confronted with the question of how to "count" the "relevant" parties. At first, this was done rather subjectively. But eventually, there was interest in developing a more abstract (and "precise") way of counting the number of parties.

The most common way to do this remains one developed by Markku Laakso and Rein Taagepera (1979), which is an inverse of the fractionalization index:

$$ENPV = \frac{1}{\sum p_i^2}$$

where p_i is the vote share (as a fraction, not a percent) of each individual party. The effective number of parties is a measure that numerically describes the number of *relevant* (or "effective") parties in a party system. Instead of ranging from zero to 1 (like the HHI and fractionalization indexes) "counts" them by giving an estimate of the number (with decimals).

We can illustrate this with an example from the 2002 election:

Table 9-3 Vote share in the 2002 Bolivian election

Party	Vote share (p_i)	Vote share squared (p_i^2)
ADN	0.0340	0.00115
CONDEPA	0.0037	0.00001
LyJ	0.0272	0.00074
MAS	0.2094	0.04385
MCPC	0.0063	0.00004
MIR	0.1632	0.02662
MIP	0.0609	0.00371
MNR	0.2246	0.05045
NFR	0.2091	0.04374
PS	0.0065	0.00004
UCS	0.0551	0.00304
$\sum p_i^2$		0.17339
$\frac{1}{\sum p_i^2}$		5.77

We convert the vote shares to fractional shares (e.g. 20% = 0.20). Then, we simply square each individual vote share, before adding them up and then diving 1 by that result. When we do that, we get a value of 5.77 "effective" parties in the 2002 Bolivian election. In other words, we can say that Bolivia was (in 2002) somewhat between a "five-party" and "six-party" system. Notice that this is smaller than the total number of parties that *competed* in the election, which was 11. The value for ENPV is intuitive, though, because we can see that four parties were relatively "equal" (MNR, MAS, MIR, and NFR) with around a fifth of the vote each, with the rest of the vote split up among several smaller parties, but most of that taken by MIP and ADN. If we look at which parties won seats, we find that only seven parties did so (and one of these, PS, only one one lonely seat in the lower house).

In the example above, we calculated the number of parties at the *national* level based on *vote shares*. We can also calculate the number of parties at lower levels (department, municipality) and we can do it with other measures, such as *seat shares*. The latter may be more appropriate if you are comparing across countries with different types of electoral systems. Some, also distinguish between the number of "legislative" parties and the number of "presidential" parties (calculating the effective number of presidential candidates).

Beyond party systems, you could also use the effective number of parties formula to "count" the "effective" number of any divisions in a society: ethnic groups, religious affiliations, occupations, etc. Again, this is a really simple formula for *transforming* or *operationalizing* variables. Just remember, as always, to avoid reification and that the indicator is only as good as the underlying data. In particular, the original Laakso and Taagpera formula has seen significant criticism because it can over/under-estimate the number of parties in circumstances where data is missing (a lot votes/seats listed for "Other" parties) or when one party is hyper-dominant. Still, there's no consensus on the "best" measure, and the Laakso and Taagepera formula remains the most widely used.

Bibliography

Alesina, Alberto, Arnaud Devleeshcauwer, William Easterly, Sergio Kurlat, Romain Wacziarg. 2003. "Fractionalization." *Journal of Economic Growth* 8: 155-194.

Baglione, Lisa A. 2016. *Writing a Research Paper in Political Science: A Practical Guide to Inquiry, Structure, and Methods*, 3rd ed. Los Angeles: Sage and CQ Press.

Dahl, Robert A. 1971. *Polyarchy: Participation and Opposition.* New Haven: Yale University Press.

Diamond, Jared. 2011. "Intra-Island and Inter-Island Comparisons." In *Natural Experiments of History*, edited by Jared Diamond and James A. Robinson. Cambridge, MA: Belknap Press of Harvard University Press.

Donovan, Todd and Kenneth Hoover. 2014. *The Elements of Social Scientific Thinking*, 11th ed. Boston: Wadsworth Publishing.

Laakso, Markku, and Rein Taagepera. 1979. "The 'Effective' Number of Parties: A Measure with Application to West Europe." *Comparative Political Studies* 12 (1): 3-27.

Lange, Matthew. 2013. *Comparative-Historical Methods*. London: Sage.

Lasswell Harold. 19848. "The Structure and Function of Communication in Society." *The Communication of Ideas* 37: 215-228.

Linz, Juan J. 1994. *The Failure of Presidential Democracy*, 2 vols. Baltimore: Johns Hopkins University Press.

Linz, Juan J. and Alfred Stepan. 1996. *Problems of Democratic Transition and Consolidation*. Baltimore: Johns Hopkins University Press.

Pedersen, Mogens. 1979. "The Dynamics of European Party Systems: Changing Patterns of Electoral Volatility" *European Journal of Political Research* 7 (1): 1-26.

Shively, W. Phillips. 2011. *The Craft of Political Research*, 8th ed. Boston: Pearson Longman.

Skocpol, Theda. 1979. *States & Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.

Teune, Henry and Adam Przeworski. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.

Thomas, Gary. 2016. *How to Do Your Case Study*, 2nd ed. London: Sage.

Vanhanen, Tatu. 1984. *The Emergence of Democracy: A Comparative Study of 119 states, 1850-1979*. Helsinki: The Finish Society of Sciences and Letters.

Wheelan, Charles. 2013. *Naked Statistics: Stripping the Dread from the Data.* New York: W. W. Norton.

Binomial Test
For nominal data. Maybe too tough to include in class?

One sample proportion test

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

will have to use NORMDIST function to figure out critical value for one tail; for two tails

This tells you the confidence interval for a value from a sample population:

$$CI = p \pm z\sqrt{\frac{p(1 - p)}{n}}$$